LINESEARCH-FREE NONLINEARLY PRECONDITIONED GD WITH SEMIAUTOMATIC GEOMETRY DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

Nonlinear preconditioning makes it possible to adapt gradient updates to the growth of the objective's curvature, but combining its design with suitable stepsize selection is challenging. We propose *adaptive dual preconditioned gradient descent* (adaptive DPGD), which combines curvature-based preconditioner design with a linesearch-free stepsize rule based on local information. For convex objectives, we establish linear convergence under local strong convexity of the objective. By adapting the stepsize rule, we extend the method to weakly convex objectives and establish asymptotic stationarity without global Lipschitz smoothness. Moreover, our local assumptions give greater freedom in preconditioner design. We exploit this freedom to develop a semiautomatic recipe guided by the objective's curvature, yielding a family of preconditioners that includes those underlying normalized and hyperbolic gradient descent. Experiments on convex and weakly convex problems with real data demonstrate the effectiveness of the preconditioner design and adaptive stepsize selection.

1 INTRODUCTION

We consider the unconstrained optimization problem

$$\min_{x\in\mathbb{R}^d} \quad f(x),$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable. The basic gradient descent (GD) iteration $x_{k+1} = x_k - \gamma_{k+1}\nabla f(x_k)$ follows the Euclidean gradient and scales it by a scalar stepsize.

Nonlinear preconditioning allows the gradient magnitude and, more generally, its direction to be adapted to the objective's curvature. Dual preconditioned gradient descent (DPGD) (Maddison et al., 2021) uses a convex preconditioning function $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, minimized at the origin, and updates

$$x_{k+1} = x_k - \gamma_{k+1}\nabla\phi(\nabla f(x_k)).$$

The choice of ϕ determines the nonlinear geometry of the update. Oikonomidis et al. (2025a;b) have shown that this framework includes updates underlying normalized gradient descent (NGD), hyperbolic gradient descent (HGD), and memoryless variants of AdaGrad (Duchi et al., 2011) and Adam (Kingma & Ba, 2015). The DPGD guarantees of Maddison et al. (2021), however, start from a prescribed preconditioner and require a global dual relative smoothness condition between f and ϕ, while providing limited guidance on how ϕ should be designed. Meanwhile, linesearch-free adaptive stepsizes have been developed for Euclidean GD using local curvature information (Malitsky & Mishchenko, 2020), with recent extensions to Bregman settings (Ou et al., 2026).

Combining these two ideas is nontrivial. Euclidean adaptive GD estimates local curvature directly from changes in gradients and iterates, whereas DPGD nonlinearly transforms the gradient through $\nabla\phi$. Its behavior therefore depends on the interaction between the curvature of f and the geometry induced by ϕ, rather than on either one alone. Consequently, a direct transfer of Euclidean adaptive rules does not provide sufficient control of the update. Our key observation is that this interaction can instead be quantified through local primal–dual Bregman curvature information. This leads to our central question:

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.

Table 1: Representative methods with convergence guarantees. NL: nonlinear preconditioning; LS: linesearch stepsize. In the Assumptions columns, Local indicates the required conditions need only hold on relevant bounded sets. cvx: convex; scvx: strongly convex; wcvx: weakly convex; ncvx: nonconvex; Dual for ∇f denotes relative smoothness/strong convexity in dual space; Anisotropic, smoothness through a preconditioner-dependent nonlinear upper bound on f.

Method	NL	Stepsize	Assumptions			Convergence
			Local	class of f	∇f smooth	
GD	✘	Fixed/LS	✘/✔	scvx	Lipschitz	Linear
Adaptive GD[a]	✘	Adaptive	✔	scvx	Lipschitz	Linear
DPGD[b]	✔	Fixed	✘	Dual scvx	Dual	Linear
DPGD-LS[b]	✔	LS	✘	cvx	Dual	$\phi(\nabla f(x_K)) - \phi(0) = O(K^{-1})$
AdaPGNC[c]	✘	Adaptive	✘	wcvx	Lipschitz	$\min_{k\leq K} \|\nabla f(x_k)\| = O(K^{-1/2})$
HGD[d]	✔	Fixed	✘	ncvx	Anisotropic	$\min_{k\leq K} \|\nabla f(x_k)\| = O(K^{-1/2})$
Adaptive DPGD	✔	Adaptive	✔	scvx	Dual	Linear
			✔	wcvx	Lipschitz	$\nabla f(x_k) \to 0$

[a]Malitsky & Mishchenko (2020); [b]Maddison et al. (2021); [c]Ye et al. (2025); [d]Oikonomidis et al. (2025a).

Can nonlinear preconditioners be designed from objective curvature while their stepsizes are adapted from local information, without global compatibility constants or repeated linesearch?

Contributions. We answer this question affirmatively by proposing *adaptive dual preconditioned gradient descent* (adaptive DPGD), which uses local Bregman curvature information for the stepsize rule. The resulting local convergence conditions provide greater freedom in the choice of ϕ, which we exploit to construct nonlinear preconditioners systematically from simple curvature models.

- **Adaptive stepsizes for DPGD.** We control the growth of the stepsize and adjust it using local curvature estimates expressed through Bregman divergences, which measure deviations from first-order approximations. The stepsize is updated directly, without a linesearch.
- **Convergence analysis.** For convex objectives, we establish linear convergence under local strong convexity of f and ϕ and local dual relative smoothness, without a global dual relative smoothness constant unlike Maddison et al. (2021). By adapting the stepsize rule to a weak convexity parameter, we preserve the DPGD update and establish $\phi(\nabla f(x_k)) \to \phi(0)$ and $\nabla f(x_k) \to 0$ without global Lipschitz smoothness.
- **Semiautomatic preconditioner design.** Local dual relative smoothness gives greater freedom in preconditioner design. We develop a semiautomatic construction using the normal and tangential inverse curvatures of radial models of nonquadratic growth. We establish that the resulting families satisfy the preconditioner assumptions of our convergence theory. We recover the preconditioners underlying normalized and hyperbolic gradient descent.

Relation to prior work Linesearch-free stepsize adaptation based on local curvature was developed for Euclidean GD by Malitsky & Mishchenko (2020) and extended to Bregman proximal-gradient methods by Ou et al. (2026). Building on the gradient-space specialization of Bregman adaPG, we establish linear convergence for adaptive DPGD under local dual relative smoothness and local strong convexity. Maddison et al. (2021) introduced DPGD under global dual relative smoothness, while Oikonomidis et al. (2025a;b) developed broader nonlinear-preconditioning frameworks under global generalized or anisotropic smoothness. For weakly convex objectives, the adaptive method of Ye et al. (2025) also requires global smoothness. Further comparisons are given in Appendix F.

2 PRELIMINARIES

Additional definitions and auxiliary results are collected in Appendix A. When a minimizer exists, write $x_\star \in \operatorname{argmin} f$ and $f_\star = f(x_\star)$. For $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, let $\operatorname{dom} h$ denote its effective domain and $\operatorname{int} \operatorname{dom} h$ its interior. For $a \in \mathbb{R}$, write $[a]_+ := \max\{a, 0\}$ and adopt the convention $c/0 := +\infty$ for $c > 0$. For open $\Omega \subset \mathbb{R}^d$ and $k \in \mathbb{N}$, $C^k(\Omega)$ denotes real-valued functions on Ω with continuous derivatives up to order k.

Definition 1 (Weak convexity). *A function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is η-weakly convex, for $\eta \geq 0$, if $x \longmapsto f(x) + \frac{\eta}{2}\|x\|^2$ is convex on $\mathbb{R}^d$ (see, e.g., Paquette et al., 2018). The case $\eta = 0$ is convexity.*

Notation 2. *Let $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be proper and lower semicontinuous, with nonempty $\Omega_h := \operatorname{int} \operatorname{dom} h$, and suppose $h \in C^1(\Omega_h)$. For $x, y \in \Omega_h$, write*

$$D_h(x, y) := h(x) - h(y) - \langle \nabla h(y), x - y \rangle, \qquad \Delta_h(x, y) := D_h(x, y) + D_h(y, x). \tag{1}$$

For $x, y, z \in \Omega_h$, the following three-point identity holds:

$$\langle \nabla h(y) - \nabla h(x), z - x \rangle = D_h(x, y) + D_h(z, x) - D_h(z, y). \tag{2}$$

We refer to D_h as the Bregman divergence generated by h and to Δ_h as its symmetrization. When h is convex, both are nonnegative.

For a differentiable $f : \mathbb{R}^d \to \mathbb{R}$, let $\mathcal{P}$ denote the class of proper, lower semicontinuous, and convex functions $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ satisfying $\nabla f(\mathbb{R}^d) \subset \Omega_\phi := \operatorname{int} \operatorname{dom} \phi$ and $\phi \in C^1(\Omega_\phi)$. The next definition localizes the global dual relative smoothness condition introduced by Maddison et al. (2021, Prop. 3.3), allowing the constant to depend on the compact set.

Definition 3 (Local dual relative smoothness). *Let $f \in C^1(\mathbb{R}^d)$ be convex and $\phi \in \mathcal{P}$. We say that ϕ is locally dual relatively smooth with respect to f if, for every nonempty compact set $\mathcal{K} \subset \mathbb{R}^d$, there exists $L_\mathcal{K} > 0$ such that*

$$D_\phi(\nabla f(x), \nabla f(y)) \leq L_\mathcal{K} D_f(y, x), \qquad x, y \in \mathcal{K}. \tag{3}$$

The corresponding local comparison bounds for the weakly convex setting, together with standard bounds on compact sets used in the convergence proofs, are collected in Appendix A.

3 ADAPTIVE DPGD ALGORITHMS

Fix a known $\eta \geq 0$ and set $g_\eta(x) := f(x) + \frac{\eta}{2}\|x\|^2$, so that $g_0 = f$. The update and stepsize rules use only function values and first derivatives. The following regularity conditions are imposed for the convergence analysis.

Assumption 4 (Objective and preconditioner).

(i) *$f \in C^2(\mathbb{R}^d)$ is level bounded, and $\nabla^2 g_\eta(x) \succ 0$ for every $x \in \mathbb{R}^d$.*

(ii) *$\phi \in \mathcal{P}$ and $\operatorname{argmin} \phi = \{0\}$.*

(iii) *If $\eta = 0$, ϕ is locally dual relatively smooth with respect to f. If $\eta > 0$, $\nabla^2 f$ and $\nabla \phi$ are locally Lipschitz on $\mathbb{R}^d$ and Ω_ϕ, respectively.*

Condition (i) gives local strong convexity of g_η and ensures that f attains its minimum; fix $x_\star \in \operatorname{argmin} f$. For $\eta = 0$, level boundedness and existence of a unique minimizer are equivalent under $\nabla^2 f \succ 0$ (Rockafellar & Wets, 1997, Thm. 2.6 and Prop. 3.23). Condition (ii) makes $\nabla \phi$ a natural preconditioner for driving ∇f toward 0. Condition (iii) retains the weaker convex regularity and ensures that $\nabla \phi \circ \nabla f$ is locally Lipschitz and $D_\phi(\nabla f(x), \nabla f(y)) \leq L_\mathcal{K} D_{g_\eta}(y, x)$ on each compact $\mathcal{K} \subset \mathbb{R}^d$. These bounds are established in Proposition 18 in Appendix A.

3.1 ADAPTIVE DPGD

Define $T_k(x) := x - \gamma_k \nabla \phi(\nabla f(x))$. For positive stepsizes, the DPGD update is

$$x_{k+1} = T_{k+1}(x_k) = x_k - \gamma_{k+1} \nabla \phi(\nabla f(x_k)), \qquad \beta_{k+1} := \frac{\gamma_{k+1}}{\gamma_k}. \tag{4}$$

By Assumption 4(ii), $x_k = x_{k-1}$ is equivalent to $\nabla f(x_{k-1}) = 0$. We terminate at stationary points and define the following quantities at nonstationary iterations, with $\delta > 0$:

$$\widehat{L}_k^{\mathrm{d},\eta} := \frac{\Delta_\phi(\nabla f(x_k), \nabla f(x_{k-1}))}{\Delta_{g_\eta}(x_k, x_{k-1})}, \qquad \mathcal{R}_{k,\delta}^{\eta} := \frac{2D_{g_\eta}(x_k + \delta[T_k(x_k) - T_k(x_{k-1})], x_k)}{\delta^2 \Delta_{g_\eta}(x_k, x_{k-1})}, \tag{5}$$

where Δ_h for some h is in (1). Set $\zeta_k^\eta := \frac{D_{g_\eta}(x_{k-1}, x_k)}{D_{g_\eta}(x_k, x_{k-1})}$ and $m_k^\eta := \frac{\Delta_{g_\eta}(x_k, x_{k-1})}{\|x_k - x_{k-1}\|^2}$. Local estimates use g_η, while updates use ∇f. The symmetrized comparison bounds $\widehat{L}_k^{\mathrm{d},\eta}$ on compact sets.

Initialize $\beta_0 = \beta_1 = 1$ and $\gamma_1 = \gamma_0$. For $\eta > 0$, fix $s \geq 0$ and $\beta_{\mathrm{cap}} \in (1, 3)$. For $k \geq 0$, define

$$\bar{\beta}_{k+1} := \begin{cases} \sqrt{1 + \beta_k}, & \text{if } \eta = 0, \\ \min\{1 + s\gamma_k, \beta_{\mathrm{cap}}\}, & \text{if } \eta > 0. \end{cases} \tag{6}$$

For $k \geq 1$, set

$$\beta_{k+1} := \begin{cases} \min\left\{\bar{\beta}_{k+1},\ \dfrac{\zeta_k^0}{1+\zeta_k^0} \dfrac{1}{2\bar{\beta}_{k+1}[\mathcal{R}_{k,2\bar{\beta}_{k+1}}^0 - (1 - \gamma_k \widehat{L}_k^{\mathrm{d},0})]_+}\right\}, & \text{if } \eta = 0, \\ \min\left\{\bar{\beta}_{k+1},\ \dfrac{\zeta_k^\eta}{1+\zeta_k^\eta} \dfrac{2(1 - \beta_k/\bar{\beta}_k) + \eta(1+\zeta_k^\eta)/(m_k^\eta \zeta_k^\eta)}{[\bar{\beta}_{k+1}\mathcal{R}_{k,\bar{\beta}_{k+1}}^\eta - 1 + \gamma_k \widehat{L}_k^{\mathrm{d},\eta} + \eta/m_k^\eta]_+}\right\}, & \text{if } \eta > 0. \end{cases} \tag{7}$$

The two cases use different Lyapunov arguments; substituting $\eta = 0$ into the positive-η formula does not recover the convex rule. When $\eta > 0$ and $s = 0$, $\bar{\beta}_{k+1} = 1$ and the stepsizes are nonincreasing. We present Adaptive DPGD with these stepsizes in Algorithm 1.

Algorithm 1: Adaptive DPGD

Input: $x_0 \in \mathbb{R}^d$, $\gamma_0 > 0$, $\eta \geq 0$; for $\eta > 0$: $s \geq 0$, $1 < \beta_{\mathrm{cap}} < 3$

Initialize $\beta_0 = \beta_1 = 1$, $\gamma_1 = \gamma_0$, and $\bar{\beta}_1$ by (6) with $k = 0$

$x_1 \leftarrow x_0 - \gamma_1 \nabla\phi(\nabla f(x_0))$

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

  Compute (5) and ζ_k^η; if $\eta > 0$, also compute m_k^η

  Set β_{k+1} using (6)–(7)

  $\gamma_{k+1} \leftarrow \beta_{k+1}\gamma_k$

  $x_{k+1} \leftarrow x_k - \gamma_{k+1}\nabla\phi(\nabla f(x_k))$

end for

4 Convergence Analysis

If $x_k = x_{k-1}$, then x_k is stationary, *i.e.*, $\nabla f(x_k) = 0$, and, when $\eta = 0$, the unique minimizer. Unless stated otherwise, we consider an infinite nonstationary sequence with $x_k \neq x_{k-1}$ for every $k \geq 1$. We present the basic identities, Lyapunov inequalities, and principal convergence results. Complete proofs are given in Appendices A to C.

We first establish two-step identities shared by the convex and weakly convex analyses. Define the gap from the optimal value of the preconditioning function by $G_k := \phi(\nabla f(x_k)) - \phi(0)$. By Assumption 4(ii), $G_k \geq 0$. Define the cross term

$$\chi_{k+1} := \beta_{k+1}\langle T_k(x_{k-1}) - T_k(x_k), \nabla f(x_k) - \nabla f(x_{k+1})\rangle. \tag{8}$$

These identities hold without convexity in both settings; see Appendix A.2 for a proof.

Lemma 5 (Basic two-step identities)**.** *Let $\{x_k\}$ be generated by (4) with positive stepsizes $\{\gamma_{k+1}\}$. Then, for every $k \geq 1$,*

$$G_{k-1} - G_k = \frac{\Delta_f(x_k, x_{k-1}) - \gamma_k \Delta_\phi(\nabla f(x_k), \nabla f(x_{k-1}))}{\gamma_k} + D_\phi(\nabla f(x_{k-1}), \nabla f(x_k)). \tag{9}$$

Furthermore, if $\bar{x} \in \mathbb{R}^d$ satisfies $\nabla f(\bar{x}) = 0$, then

$$D_f(x_{k+1}, \bar{x}) + \gamma_{k+1}G_k + D_f(x_k, x_{k+1}) = D_f(x_k, \bar{x}) + \chi_{k+1} - \gamma_{k+1}D_\phi(0, \nabla f(x_k)). \tag{10}$$

4.1 CONVEX OBJECTIVES

Throughout this subsection, $\eta = 0$; we omit the superscript 0 from the local quantities. We combine the two identities in Lemma 5 and bound the cross term χ_{k+1} using the Bregman three-point identity (2). This yields a decrease inequality for the Lyapunov function $\mathcal{E}_k$ defined below. The proof is given in Appendix B.1.

Lemma 6 (Lyapunov inequality)**.** *Define the Lyapunov function by*

$$\mathcal{E}_k := D_f(x_k, x_\star) + \gamma_k(1 + \bar{\beta}_k)G_{k-1} + \left(1 - \frac{\beta_k}{2\bar{\beta}_k}\right) D_f(x_{k-1}, x_k). \tag{11}$$

Under the update rule in Algorithm 1 and Assumption 4 with $\eta = 0$, for every $k \geq 1$,

$$\mathcal{E}_{k+1} \leq \mathcal{E}_k - \gamma_k(1 + \bar{\beta}_k - \beta_{k+1}\bar{\beta}_{k+1})G_{k-1} \leq \mathcal{E}_k.$$

This Lyapunov inequality yields the following main result.

Theorem 7. *Suppose that Assumption 4 holds with $\eta = 0$. Then the sequence $\{x_k\}$ generated by Algorithm 1 satisfies $x_k \to x_\star$ and $\nabla f(x_k) \to 0$.*

Proof sketch of Theorem 7. Lemma 6 gives boundedness of the iterates. On a compact set containing the sequence, $\widehat{L}_k^{\mathrm{d}}$ is uniformly bounded above, ζ_k is bounded away from 0, and $\mathcal{R}_{k,2\bar{\beta}_{k+1}} \to 1$ if $\gamma_k \to 0$. These local estimates and the stepsize rule imply $\sum_k \gamma_k = \infty$. Combining the existence of a subsequence approaching stationarity with an argument about the limit of the Lyapunov function proves convergence of the entire sequence to $x_\star$. The complete proof is given in Appendix B.5. □

For the convergence rate, we make an additional assumption.

Assumption 8. *For every nonempty compact convex set $\mathcal{K}^* \subset \Omega_\phi$, there exists $\mu_{\mathcal{K}^*} > 0$ such that ϕ is $\mu_{\mathcal{K}^*}$-strongly convex on $\mathcal{K}^*$.*

We will discuss the construction of such preconditioners in Section 5, which shows how to construct preconditioners satisfying Assumption 8.

Theorem 9 (R-linear convergence)**.** *Suppose that Assumptions 4 and 8 hold with $\eta = 0$. Then there exist $C_f, C_\nabla > 0$ and $\alpha \in (0, 1)$ such that the sequence generated by Algorithm 1 with the convex stepsize rule satisfies, for all $k \geq 0$,*

$$0 \leq f(x_k) - f_\star \leq C_f \alpha^k, \qquad \|\nabla f(x_k)\| \leq C_\nabla \alpha^k. \tag{12}$$

Proof sketch of Theorem 9. We do not require a positive lower bound on the one-step Lyapunov decrease coefficient. Local bounds and Assumption 8 bound the stepsizes above and away from zero and give $\mathcal{E}_k \leq C_E G_{k-1}$ for some $C_E > 0$. Summing over a fixed number N of iterations yields $\mathcal{E}_{k+N} \leq \kappa \mathcal{E}_k$ with $\kappa \in (0, 1)$, implying (12). The complete proof is in Appendix B.6. □

4.2 WEAKLY CONVEX OBJECTIVES

Throughout this subsection, Assumption 4 holds with $\eta > 0$. We use the local comparison bound (28), established in Proposition 18 in Appendix A, with the convexified function g_η. Since $x_\star$ globally minimizes f on $\mathbb{R}^d$, $\nabla f(x_\star) = 0$ and

$$D_f(x_k, x_\star) = f(x_k) - f(x_\star) \geq 0. \tag{13}$$

Moreover, the definition of g_η gives

$$\Delta_f(x_k, x_{k-1}) = \Delta_{g_\eta}(x_k, x_{k-1}) - \eta\|x_k - x_{k-1}\|^2$$

and, using (5) and m_k^η,

$$\Delta_f(x_k, x_{k-1}) - \gamma_k \Delta_\phi(\nabla f(x_k), \nabla f(x_{k-1})) = \left(1 - \gamma_k \widehat{L}_k^{\mathrm{d},\eta} - \frac{\eta}{m_k^\eta}\right) \Delta_{g_\eta}(x_k, x_{k-1}). \tag{14}$$

We next establish the Lyapunov inequality for weakly convex objectives. The proof is given in Appendix C.1.

Lemma 10 (Lyapunov inequality)**.** *Define the Lyapunov function by*

$$\mathcal{E}_k^{\mathrm{weak}} := D_f(x_k, x_\star) + \frac{3}{2}\gamma_k G_{k-1} + \left(1 - \frac{\beta_k}{\bar{\beta}_k}\right) D_{g_\eta}(x_{k-1}, x_k) + \frac{\eta}{2}\|x_k - x_{k-1}\|^2. \tag{15}$$

Under Algorithm 1 with $\eta > 0$, for every $k \geq 1$,

$$\mathcal{E}_{k+1}^{\mathrm{weak}} \leq \mathcal{E}_k^{\mathrm{weak}} - \frac{\gamma_k}{2}(3 - \beta_{k+1})G_{k-1}. \tag{16}$$

Since $\beta_{k+1} \leq \beta_{\mathrm{cap}} < 3$, the coefficient of $\gamma_k G_{k-1}$ in the decrease is uniformly positive.

The Lyapunov inequality yields the following stationarity result.

Theorem 11 (Stationarity for weakly convex objectives)**.** *Under Assumption 4 with $\eta > 0$ and Algorithm 1,*

$$\phi(\nabla f(x_k)) \to \phi(0). \qquad \textit{Consequently,} \quad \nabla f(x_k) \to 0. \tag{17}$$

Proof sketch of Theorem 11. Lemma 10 gives boundedness of the iterates. If $\gamma_k \to 0$, local estimates yield $\gamma_{k+1} \geq \gamma_k(1 - C\gamma_k)$ eventually for some $C > 0$, so $\sum_k \gamma_k = \infty$ in all cases. Combining this with $\sum_k \gamma_k G_{k-1} < \infty$ gives $\liminf_k G_k = 0$. Local Lipschitz continuity of $\phi \circ \nabla f$ then yields $G_k \to 0$ and (17). The complete proof is in Appendix C.5. □

5 Semiautomatic Design of Preconditioners

We construct nonlinear preconditioners from simple curvature models of the objective. Unlike fixed-stepsize DPGD analyses (Maddison et al., 2021) that require a problem-specific global compatibility condition for a prescribed preconditioner, our local framework allows us to design ϕ from the curvature behavior that we wish to correct. The construction is semiautomatic: once a target model, a curvature direction, and an interpolation parameter are selected, the preconditioner is determined.

Design principle. For convex f, the ideal relation $\nabla^2\phi(\nabla f(x)) \approx (\nabla^2 f(x))^{-1}$ would make the Jacobian of the preconditioned direction, $\nabla^2\phi(\nabla f(x))\nabla^2 f(x)$, close to the identity. We therefore design radial preconditioners that nonlinearly rescale gradients to compensate for nonquadratic curvature growth. We model the growth to be corrected by a radial target $t(x) = \tau(\|x\|)$, where $\tau : [0, \infty) \to \mathbb{R}$ is C^2 on $(0, \infty)$ with $\tau'(r) > 0$ and $\tau''(r) > 0$ for $r > 0$. For $r = \|x\| > 0$, $\nabla^2 t(x) = \tau''(r)\frac{xx^\top}{r^2} + \frac{\tau'(r)}{r}\left(I - \frac{xx^\top}{r^2}\right)$. Thus the normal and tangential curvatures are $\tau''(r)$ and $\tau'(r)/r$, respectively. We select the corresponding inverse curvature $h_{\mathrm{N}} = (\tau''(r))^{-1}$ or $h_{\mathrm{T}} = (\tau'(r)/r)^{-1}$ and express it as a function of the gradient magnitude $\varrho := \|\nabla t(x)\|$.

Interpolation and preconditioner families. Suppose that the selected inverse curvature, expressed as a function of ϱ, has a leading power-law factor ϱ^a for some $a \in \mathbb{R}$. The exponent a captures the asymptotic inverse-curvature growth of the target model. To preserve unit curvature near the origin while matching this behavior for large gradients, define, for $b > 0$,

$$\psi_{a,b}(\varrho) := (1 + \varrho^b)^{a/b}, \qquad \varrho \geq 0. \tag{18}$$

Then $\psi_{a,b}(0) = 1$ and $\psi_{a,b}(\varrho)/\varrho^a \to 1$ as $\varrho \to \infty$. Hence the resulting geometry is GD-like for small gradients and matches the selected power-law factor for large gradients. The exponent a is determined by the target curvature, whereas b controls the transition.

We next construct a radial preconditioner $\phi(y) = \nu(\|y\|)$ whose normal or tangential curvature matches $\psi_{a,b}$. Since these curvatures are $\nu''(\varrho)$ and $\nu'(\varrho)/\varrho$, respectively, integration gives

$$\nu_{\mathrm{N}}^{a,b}(\varrho) := \int_0^\varrho (\varrho - s)\psi_{a,b}(s)\, ds, \qquad \nu_{\mathrm{T}}^{a,b}(\varrho) := \int_0^\varrho s\psi_{a,b}(s)\, ds, \tag{19}$$

and we set $\phi_{\mathrm{N}}^{a,b}(y) := \nu_{\mathrm{N}}^{a,b}(\|y\|)$, $\phi_{\mathrm{T}}^{a,b}(y) := \nu_{\mathrm{T}}^{a,b}(\|y\|)$. Thus $\phi_{\mathrm{N}}^{a,b}$ matches $\psi_{a,b}$ in its normal curvature, while $\phi_{\mathrm{T}}^{a,b}$ matches it in its tangential curvature. For the tangential family $\phi_{\mathrm{T}}^{a,b}$, we restrict to $a \geq -1$.

For weakly convex objectives, the Hessian of f need not be positive definite, so the inverse-Hessian interpretation above is no longer literal. We instead use the radial target as a model of the curvature behavior to be corrected, while applying the same preconditioner construction.

Table 2: Preconditioning functions and related methods, where $q = p/(p-1) > 1$.

Preconditioning function	Relation
$\phi_{\mathrm{N}}^{0,b}(y) = \phi_{\mathrm{T}}^{0,b}(y) = \dfrac{\|y\|^2}{2} \quad (b > 0)$	GD
$\phi_{\mathrm{N}}^{q-2,1}(y) = \dfrac{(1+\|y\|)^q - 1 - q\|y\|}{q(q-1)}$	New variant
$\phi_{\mathrm{T}}^{q-2,2}(y) = \dfrac{(1+\|y\|^2)^{q/2} - 1}{q}$	p-norm DPGD (Maddison et al., 2021)
$\phi_{\mathrm{N}}^{-1,1}(y) = (1+\|y\|)\log(1+\|y\|) - \|y\|$	Oikonomidis et al. (2025a, Table 1)
$\phi_{\mathrm{N}}^{-1,2}(y) = \|y\|\operatorname{arsinh}(\|y\|) - \sqrt{1+\|y\|^2} + 1$	HGD (Oikonomidis et al., 2025a;b)
$\phi_{\mathrm{T}}^{-1,1}(y) = \phi_{\mathrm{N}}^{-2,1}(y) = \|y\| - \log(1+\|y\|)$	NGD; exponential-penalty DPGD (Zhang et al., 2020; Maddison et al., 2021)
$\phi_{\mathrm{T}}^{-1,2}(y) = \sqrt{1+\|y\|^2} - 1$	AdaGrad-type (Duchi et al., 2011)
$\lim_{b\to\infty} \phi_{\mathrm{T}}^{-1,b}(y) = \begin{cases} \|y\|^2/2, & \|y\| \le 1, \\ \|y\| - 1/2, & \|y\| > 1 \end{cases}$	Gradient clipping (Zhang et al., 2020)

Choosing the exponent from the target model. The exponent a follows directly from the target curvature. For the power-law target $\tau(r) = r^p/p$ with $p > 1$, let $q := p/(p-1)$. Since $\varrho = \tau'(r) = r^{p-1}$, both inverse curvatures are proportional to ϱ^{q-2}, giving $a = q - 2 = (2-p)/(p-1)$; for example, $p = 6$ gives $a = -0.8$. For the exponential target $\tau(r) = e^r$, we have $h_{\mathrm{N}}(\varrho) = \varrho^{-1}$. Thus the normal construction gives $a = -1$. Further details are given in Appendix D.2.

Table 2 lists selected preconditioning functions and related methods. Further examples are given in Appendix D.3. The family also highlights different large-gradient behaviors. The AdaGrad-type member $\phi_{\mathrm{T}}^{-1,2}$ asymptotically normalizes the gradient, whereas the new power-law variant $\phi_{\mathrm{N}}^{q-2,1}$ produces a preconditioned gradient with norm of order $\|y\|^{q-1}$, with $q - 1 = 1/(p-1)$. Thus the latter adapts the scaling of the preconditioned gradient to the curvature growth of the target model.

We finally verify that the constructed preconditioner families satisfy the regularity conditions required by the convergence analyses in Sections 3 and 4. We omit the superscripts when a, b are fixed. The proof is given in Appendix D.1.

Proposition 12. *Let $a \in \mathbb{R}$ and $b > 0$, and let $\phi = \phi_{\mathrm{N}}$ or, when $a \ge -1$, $\phi = \phi_{\mathrm{T}}$. Then:*

(i) *$\phi \in C^2(\mathbb{R}^d)$, $\Omega_\phi = \mathbb{R}^d$, $\nabla^2\phi \succ 0$, $\nabla^2\phi(0) = I$, and $\operatorname{argmin}\phi = \{0\}$. Thus Assumption 4(ii) holds and $\nabla\phi$ is locally Lipschitz.*

(ii) *If f satisfies Assumption 4(i) with $\eta = 0$, then Assumptions 4 and 8 hold.*

(iii) *If f satisfies Assumption 4(i) with $\eta > 0$ and $\nabla^2 f$ is locally Lipschitz, Assumption 4 holds.*

6 Numerical Experiments

We test curvature-based design and adaptive stepsizes on convex and weakly convex problems. We use Adaptive DPGD with $\phi_{\mathrm{N}}^{a,1}$ (ours), which performed well across the tested problems, with a determined by each target. We use a linesearch only to initialize the first stepsize for Adaptive DPGD; all subsequent steps are linesearch-free. We compare it with Euclidean Adaptive DPGD, with $\phi(y) = \|y\|^2/2$, to assess the effect of nonlinear preconditioning, and with GD Armijo (Armijo, 1966) and, for convex problems, adaptive GD of Malitsky & Mishchenko (2020). Additional preconditioner examples and convergence comparisons are given in Appendix E.4. We use adaptive GD with $\lambda_0 = 10^{-10}$ and $\theta_0 = +\infty$ as in Malitsky & Mishchenko (2020), and weakly convex

Adaptive DPGD with $s = 0.1$ and $\beta_{\text{cap}} = 2$. Linesearches use dyadic trials, with Armijo coefficient 10^{-4}. Initial line-search candidates are 1 for ℓ_p loss and weakly convex problems. For powered hinge loss, the initial inverse stepsize for DPGD is $\|\nabla\phi(\nabla f(x_0))\|$, and the initial Armijo stepsize is $1/\|\nabla f(x_0)\|$.

Figure 1 plots the relative objective gap $(f(x_k) - f_{\text{ref}})/(f(x_0) - f_{\text{ref}})$ for convex problems and the relative gradient norm $\|\nabla f(x_k)\|/\|\nabla f(x_0)\|$ for weakly convex problems, against runtime. Reference values, timing, and the computational environment are detailed in Appendix E.3. Runs stop when the relative gradient norm is at most 10^{-6} for five consecutive iterations. For the convex comparisons in Figure 1, we use nonzero initial points to examine convergence from regions of pronounced nonquadratic curvature growth. Preprocessing and initial points are given in Appendix E.1. Full plots, including gradient evaluations, are in Appendix E.5.

Convex problems. For data rows $b_i^\top$ of $B \in \mathbb{R}^{n\times d}$, responses $y \in \mathbb{R}^n$, and $p > 2$, we use the regularized ℓ_p loss: $\min_{x\in\mathbb{R}^d} \frac{1}{np}\sum_{i=1}^n |b_i^\top x - y_i|^p + \frac{1}{2n}\|x\|^2$. As in DPGD for p-norm regression (Maddison et al., 2021), the quadratic term ensures a positive definite Hessian. We use $\tau(r) = r^p/p$ (Appendix D.2). We set $p = 6$ on `covtype` (Blackard, 1998; Chang & Lin, 2011) ($n = 581012$, $d = 53$). Our method reduces the objective gap faster than the three baselines (Figure 1(a)).

For binary labels $y_i \in \{-1, 1\}$ and feature rows $a_i^\top$, we also consider powered hinge loss (Chen et al., 2004): $\min_{x\in\mathbb{R}^d} \frac{1}{np}\sum_{i=1}^n [1 - y_i a_i^\top x]_+^p + \frac{1}{2n}\|x\|^2$, with $p = 6$ and the target $\tau(r) = r^p/p$. On `a9a` (Chang & Lin, 2011; Becker & Kohavi, 1996) ($n = 32561$, $d = 123$), our method reaches a relative objective gap of 10^{-6} sooner than the three baselines (Figure 1(b)). We use high-dimensional TF–IDF data for `LEDGAR` ($n = 70000$, $d = 19996$) and `SCOTUS` ($n = 6400$, $d = 126405$), distributed by LIBSVM (Chang & Lin, 2011), with the most frequent class versus the rest. Our method gives a rapid initial reduction of the objective gap and reaches the gradient stopping criterion sooner than the three baselines on both datasets (Figure 1(c),(d)).

Weakly convex problems. For data rows $a_i^\top$ of $A \in \mathbb{R}^{n\times d}$ and labels $y \in \{-1, 1\}^n$, we combine exponential loss (Friedman et al., 2000) and a Cauchy penalty (Karakuş et al., 2020): $\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \exp(-y_i a_i^\top x) + \omega\sum_{j\in\mathcal{J}} \log(1 + x_j^2/\sigma^2)$. We exclude the intercept from $\mathcal{J}$, set $\omega = 0.1$, $\sigma = 1$, and use $\eta = \omega/(4\sigma^2)$ and $\tau(r) = \exp(r)$. The global weak convexity bound is derived in Appendix E.2. On `Skin Non-Skin` (Bhatt & Dhall, 2009) ($n = 245057$, $d = 4$) and `Rice Cammeo/Osmancik` (Cinar & Koklu, 2019) ($n = 3810$, $d = 8$), our method reduces the gradient norm faster than its Euclidean variant and GD Armijo (Figure 1(e),(f)). Across the three methods, terminal objective values agree within 5×10^{-12} on both datasets.

Effects of adaptive stepsizes and preconditioner design. The normal constructions provide effective alternatives to existing preconditioners, with competitive runtime even for the numerically evaluated $\phi_{\text{N}}^{a,2}$ (Appendix E.4.4). On `covtype`, the powered hinge datasets, and the weakly convex problems, Adaptive DPGD reaches the stopping criterion with fewer gradient evaluations and less runtime than its linesearch counterparts in nearly all comparisons, illustrating the benefit of avoiding repeated trial evaluations. For example, on `covtype` with $\phi_{\text{N}}^{a,1}$, Adaptive DPGD uses 94 gradient evaluations and 2.25 seconds, versus 194 evaluations and 3.10 seconds for DPGD linesearch.

Varying only a in $\phi_{\text{N}}^{a,1}$, the model-derived exponents $a = -0.8$ on `svmguide1` and $a = -1$ on `Skin Non-Skin` give faster convergence than the alternative exponents in Figure 2. This directly supports choosing a from the curvature model rather than treating it as a tunable hyperparameter.

7 Conclusion

We proposed Adaptive DPGD, combining curvature-based nonlinear preconditioner design with linesearch-free stepsizes. Without specifying a global dual relative smoothness constant, we prove linear convergence of the objective gap and gradient norm under local strong convexity in convex problems, and stationarity with the same explicit update in weakly convex problems. Real-data experiments demonstrate the benefits of both preconditioner design and adaptive stepsizes.

The framework suggests several natural extensions. For weakly convex problems, it would be interesting to obtain quantitative stationarity rates and to adapt to an unknown weak-convexity parameter

while retaining the explicit DPGD update. On the design side, the normal/tangential choice and interpolation parameter could be selected automatically from objective or iterate information. The same ideas also motivate coordinatewise or blockwise curvature estimates for separable preconditioners related to AdaGrad and SNGD/SHGD (Duchi et al., 2011; Zhang et al., 2020; Oikonomidis et al., 2025a;b) as discussed in Appendix D.3.

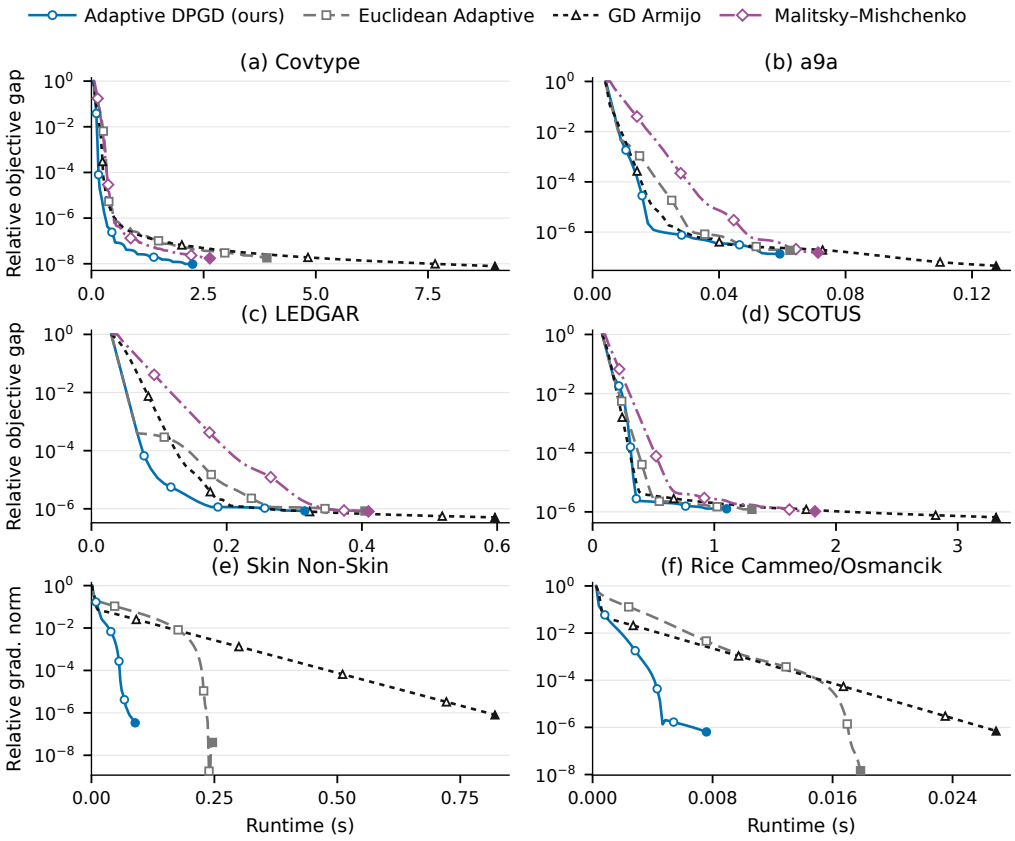


Figure 1: Runtime comparisons. Adaptive DPGD (ours) uses $\phi_{\mathrm{N}}^{a,1}$. (a) ℓ_6 loss on `covtype`; (b–d) powered hinge loss ($p = 6$) on `a9a`, `LEDGAR`, and `SCOTUS`; (e–f) exponential loss with Cauchy penalty on `Skin Non-Skin` and `Rice Cammeo/Osmancik`. Convex panels show relative objective gaps; weakly convex panels show relative gradient norms.

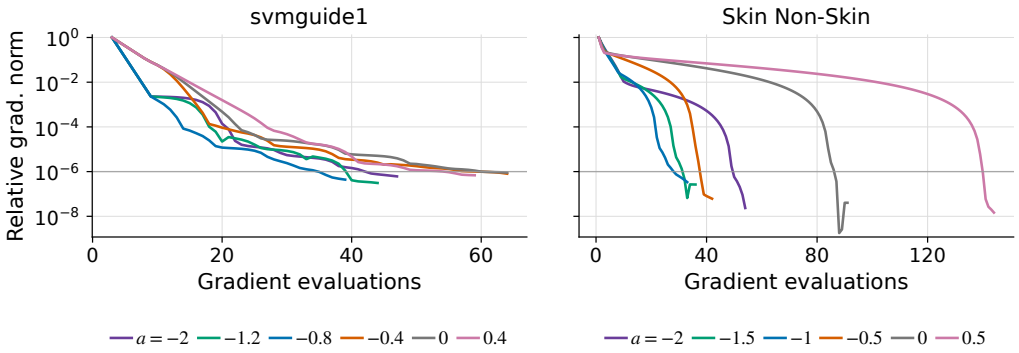


Figure 2: Adaptive DPGD with $\phi_{\mathrm{N}}^{a,1}$, varying only a. Model exponents: -0.8 (powered hinge loss, $p = 6$, `svmguide1`) and -1 (`Skin Non-Skin`); $a = 0$ is Euclidean Adaptive DPGD.

AI USE STATEMENT

In this work, we used generative AI tools for developing theoretical frameworks, formulating mathematical claims, providing key proof ingredients, assisting with proof writing, proposing or refining hypotheses, providing feedback on research methodology and experiments, implementing methods, assisting with translation, writing data-preprocessing code, and interpreting results. We have not used generative AI tools for generating synthetic datasets, and qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools for preparing scientific figures, suggesting experimental parameters, writing or editing code, drafting and editing manuscript text, searching and summarizing literature, and formatting references. We have reviewed all AI-assisted work. The authors checked the proofs and manuscript text and executed the experimental code to inspect its behavior and outputs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work concerns the development and theoretical analysis of optimization algorithms, with numerical experiments on existing publicly available benchmark datasets. We do not identify any specific ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

The algorithms, assumptions, and preconditioner constructions are described in Sections 3 to 5, with proofs in Appendices A to D. The experimental settings are described in Section 6, with datasets, preprocessing, and initial points in Appendix E.1 and reference values, timing, and the computational environment in Appendix E.3. Code for the numerical experiments and data preprocessing is provided as supplementary material.

REFERENCES

Larry Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3, 1966.

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.

Amir Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996.

Rajen Bhatt and Abhinav Dhall. Skin Segmentation. UCI Machine Learning Repository, 2009.

Jock Blackard. Covertype. UCI Machine Learning Repository, 1998.

Jérôme Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd. First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4310–4330. Association for Computational Linguistics, 2022.

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

Di-Rong Chen, Qiang Wu, Yiming Ying, and Ding-Xuan Zhou. Support vector machine soft margin classifiers: Error analysis. *Journal of Machine Learning Research*, 5:1143–1175, 2004.

Ilkay Cinar and Murat Koklu. Classification of rice varieties using artificial intelligence methods. *International Journal of Intelligent Systems and Applications in Engineering*, 7(3):188–194, 2019.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2): 337–407, 2000.

Kiwamu Fujiki, Shota Takahashi, and Akiko Takeda. Approximate Bregman proximal gradient algorithm with variable metric Armijo–Wolfe line search. *arXiv preprint arXiv:2510.06615*, 2025.

Pham Thi Hoai and Nguyen Pham Duy Thai. Composite optimization problems via novel proximal gradient algorithms and applications. Optimization Online, 2024.

Oktay Karakuş, Perla Mayo, and Alin Achim. Convergence guarantees for non-convex optimisation with Cauchy-based penalties. *IEEE Transactions on Signal Processing*, 68:6159–6170, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Tianjiao Li and Guanghui Lan. A simple uniformly optimal method without line search for convex optimization. *Mathematical Programming*, 219(1):219–256, 2026.

Chris J. Maddison, Daniel Paulin, Yee Whye Teh, and Arnaud Doucet. Dual space preconditioning for gradient descent. *SIAM Journal on Optimization*, 31(1):991–1016, 2021.

Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6702–6712. PMLR, 2020.

Mahesh Chandra Mukkamala and Peter Ochs. Beyond alternating updates for matrix factorization with inertial Bregman proximal gradient algorithms. In *Advances in Neural Information Processing Systems*, volume 32, pp. 4268–4278. Curran Associates, Inc., 2019.

Arkadi S. Nemirovski and David B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Series in Discrete Mathematics. Wiley, Chichester, 1983.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004.

Konstantinos Oikonomidis, Jan Quan, Emanuel Laude, and Panagiotis Patrinos. Nonlinearly preconditioned gradient methods under generalized smoothness. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47132–47154. PMLR, 2025a.

Konstantinos Oikonomidis, Jan Quan, and Panagiotis Patrinos. Nonlinearly preconditioned gradient methods: Momentum and stochastic analysis. In *Advances in Neural Information Processing Systems*, volume 38, pp. 38957–38988. Curran Associates, Inc., 2025b.

Hongjia Ou, Puya Latafat, and Andreas Themelis. Linesearch-free adaptive Bregman proximal gradient for convex minimization under local relative smoothness. *Journal of Optimization Theory and Applications*, 210(3):68, 2026.

Courtney Paquette, Hongzhou Lin, Dmitriy Drusvyatskiy, Julien Mairal, and Zaid Harchaoui. Catalyst for gradient-based nonconvex optimization. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 613–622. PMLR, 2018.

Jiyoung Park, Jaewook J. Suh, Bofan Wang, Anirban Bhattacharya, and Shiqian Ma. Adaptive gradient descent on Riemannian manifolds and its applications to Gaussian variational inference. In *International Conference on Learning Representations*, 2026.

Vivak Patel and Albert S. Berahas. Gradient descent in the absence of global Lipschitz continuity of the gradients. *SIAM Journal on Mathematics of Data Science*, 6(3):602–626, 2024.

Ralph Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, Germany, 1st edition, 1997.

Jaewook J. Suh and Shiqian Ma. An adaptive and parameter-free Nesterov's accelerated gradient method for convex optimization. *arXiv preprint arXiv:2505.11670*, 2025.

Shota Takahashi. Adaptive conditional gradient sliding: Projection-free and line-search-free acceleration. *arXiv preprint arXiv:2601.20443*, 2026.

Shota Takahashi and Akiko Takeda. Approximate Bregman proximal gradient algorithm for relatively smooth nonconvex optimization. *Computational Optimization and Applications*, 90(1): 227–256, 2025.

Shota Takahashi, Mituhiro Fukuda, and Mirai Tanaka. New Bregman proximal type algorithms for solving DC optimization problems. *Computational Optimization and Applications*, 83(3): 893–931, 2022.

Shota Takahashi, Mirai Tanaka, and Shiro Ikeda. Majorization-minimization Bregman proximal gradient algorithms for NMF with the Kullback–Leibler divergence. *Journal of Optimization Theory and Applications*, 208(1):14, 2025.

Shota Takahashi, Sebastian Pokutta, and Akiko Takeda. Fast Frank–Wolfe algorithms with adaptive Bregman step-size for weakly convex functions. In *International Conference on Learning Representations*, 2026.

Don Tuggener, Pius von Däniken, Thomas Peetz, and Mark Cieliebak. LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 1235–1241. European Language Resources Association, 2020.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15:49–60, 2013.

Zeyi Xu and Long Chen. Adaptive accelerated mirror descent in primal and dual spaces. *arXiv preprint arXiv:2606.02787*, 2026.

Shotaro Yagishita and Masaru Ito. Simple linesearch-free first-order methods for nonconvex optimization. *arXiv preprint arXiv:2509.14670*, 2025.

Zilong Ye, Shiqian Ma, Junfeng Yang, and Danqing Zhou. A simple adaptive proximal gradient method for nonconvex optimization. *arXiv preprint arXiv:2510.06079*, 2025.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020.

A Technical Preliminaries

We collect local regularity definitions, auxiliary results on compact sets, and the proof of the two-step identities used in both convergence analyses.

A.1 Definitions and preliminary results

Definition 13 (Level boundedness)**.** *A function h is level bounded if, for every $c \in \mathbb{R}$,*

$$\{x \in \mathbb{R}^d \mid h(x) \leq c\}$$

is bounded; see, e.g., Rockafellar & Wets (1997, Definition 1.8).

Definition 14 (Local Lipschitz continuity)**.** *A map S on an open set $\Omega \subset \mathbb{R}^d$ is locally Lipschitz continuous if, for each $x \in \Omega$, there exist a neighborhood $U_x \subset \Omega$ and a constant $L_x > 0$ such that, for every $y, z \in U_x$,*

$$\|S(y) - S(z)\| \le L_x \|y - z\|.$$

The pointwise property is called strict continuity in Rockafellar & Wets (1997, Definition 9.1(b)).

Definition 15 (Strong convexity on a set and local strong convexity)**.** *A differentiable convex function h is $\mu_{\mathcal{K}}$-strongly convex on a convex set $\mathcal{K} \subset \operatorname{int}\operatorname{dom} h$ if $\mu_{\mathcal{K}} > 0$ and, for all $x, y \in \mathcal{K}$,*

$$h(y) \ge h(x) + \langle \nabla h(x), y - x \rangle + \frac{\mu_{\mathcal{K}}}{2} \|y - x\|^2. \tag{20}$$

The first-order condition above is the standard characterization of strong convexity restricted to $\mathcal{K}$; see, e.g., Beck (2017, Theorem 5.24). We call h locally strongly convex if it is strongly convex on every nonempty compact convex set $\mathcal{K} \subset \operatorname{int}\operatorname{dom} h$.

Lemma 16. *Let $h \in C^2(\Omega)$ be convex, with $\nabla^2 h(x) \succ 0$ for all $x \in \Omega$. Then h is locally strongly convex on Ω. In particular, for every nonempty compact convex set $\mathcal{K} \subset \Omega$, there exist $m_{\mathcal{K}}, M_{\mathcal{K}} > 0$ such that*

$$m_{\mathcal{K}} I \preceq \nabla^2 h(x) \preceq M_{\mathcal{K}} I, \qquad x \in \mathcal{K}. \tag{21}$$

Furthermore, for every $x, y \in \mathcal{K}$,

$$\frac{m_{\mathcal{K}}}{2} \|x - y\|^2 \le D_h(x, y) \le \frac{M_{\mathcal{K}}}{2} \|x - y\|^2 \tag{22}$$

and

$$m_{\mathcal{K}} \|x - y\|^2 \le \Delta_h(x, y) \le M_{\mathcal{K}} \|x - y\|^2, \qquad m_{\mathcal{K}} \|x - y\| \le \|\nabla h(x) - \nabla h(y)\| \le M_{\mathcal{K}} \|x - y\|. \tag{23}$$

In particular, if $x \ne y$, then

$$\frac{m_{\mathcal{K}}}{M_{\mathcal{K}}} \le \frac{D_h(x, y)}{D_h(y, x)} \le \frac{M_{\mathcal{K}}}{m_{\mathcal{K}}}. \tag{24}$$

Proof of Lemma 16. Fix a nonempty compact convex set $\mathcal{K} \subset \Omega$. Continuity of $\nabla^2 h$ and compactness of $\mathcal{K}$ imply that $x \mapsto \lambda_{\min}(\nabla^2 h(x))$ has a positive minimum on $\mathcal{K}$, and $x \mapsto \lambda_{\max}(\nabla^2 h(x))$ has a finite maximum. Denoting these by $m_{\mathcal{K}}, M_{\mathcal{K}} > 0$ gives (21).

Fix $x, y \in \mathcal{K}$ and $w := x - y$. Convexity of $\mathcal{K}$ gives $y + tw \in \mathcal{K}$ for $0 \le t \le 1$, and Taylor's integral formula yields

$$D_h(x, y) = \int_0^1 (1 - t) \langle \nabla^2 h(y + tw) w, w \rangle \, dt.$$

Integrating (21) gives (22). Similarly,

$$\Delta_h(x, y) = \langle \nabla h(x) - \nabla h(y), w \rangle = \int_0^1 \langle \nabla^2 h(y + tw) w, w \rangle \, dt$$

gives the bounds on Δ_h in (23). Moreover, the argument of Beck (2017, Theorem 5.12), applied along segments in $\mathcal{K}$, uses

$$\nabla h(x) - \nabla h(y) = \int_0^1 \nabla^2 h(y + tw) w \, dt$$

together with (21) to obtain $\|\nabla h(x) - \nabla h(y)\| \le M_{\mathcal{K}} \|w\|$. Conversely, Cauchy–Schwarz and $\Delta_h(x, y) \ge m_{\mathcal{K}} \|w\|^2$ give $\|\nabla h(x) - \nabla h(y)\| \ge m_{\mathcal{K}} \|w\|$. Equation (20) is equivalent to the lower bound in (22). Finally, for $x \ne y$, apply (22) to (x, y) and (y, x) and take ratios to obtain (24). □

The next lemma gives local Lipschitz continuity of the composition used in the convex convergence analysis.

Lemma 17. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be C^2, with $\nabla^2 f(x) \succ 0$ for every $x \in \mathbb{R}^d$. Suppose that $\phi \in \mathcal{P}$ is locally dual relatively smooth with respect to f. Then $\nabla \phi \circ \nabla f$ is locally Lipschitz continuous on $\mathbb{R}^d$.*

Proof of Lemma 17. Fix $\bar{x} \in \mathbb{R}^d$. Positive definiteness of $\nabla^2 f(\bar{x})$, continuity of $\nabla^2 f$, and the inverse function theorem (Rockafellar & Wets, 1997, Corollary 9.55) provide a bounded open convex neighborhood $\mathcal{V}$ of $\bar{x}$, an open convex neighborhood $\mathcal{W}$ of $\nabla f(\bar{x})$, and constants $m, M > 0$ such that

$$\overline{\mathcal{V}} \text{ is compact}, \qquad \mathcal{W} \subset \nabla f(\mathcal{V}) \cap \Omega_\phi$$

and

$$m\|x - y\| \le \|\nabla f(x) - \nabla f(y)\| \le M\|x - y\|, \qquad x, y \in \mathcal{V}. \tag{25}$$

Moreover, $\nabla^2 f(x) \preceq MI$ on $\mathcal{V}$ gives

$$D_f(y, x) \le \frac{M}{2}\|y - x\|^2, \qquad x, y \in \mathcal{V}. \tag{26}$$

For $u, v \in \mathcal{W}$, set

$$x := (\nabla f|_{\mathcal{V}})^{-1}(u), \qquad y := (\nabla f|_{\mathcal{V}})^{-1}(v).$$

By (3), (25), and (26),

$$D_\phi(u, v) \le L_{\overline{\mathcal{V}}} D_f(y, x) \le \frac{L_{\overline{\mathcal{V}}} M}{2}\|y - x\|^2 \le \frac{L_{\overline{\mathcal{V}}} M}{2m^2}\|u - v\|^2. \tag{27}$$

Set $C := L_{\overline{\mathcal{V}}} M/m^2 > 0$. We localize the argument of Nesterov (2004, Theorem 2.1.5). Continuity of $\nabla\phi$ allows us to choose an open ball $\mathcal{U}$ about $\nabla f(\bar{x})$, contained in $\mathcal{W}$, such that, for all $u, v \in \mathcal{U}$, the points $v - \frac{\nabla\phi(v) - \nabla\phi(u)}{C}$ lie in $\mathcal{W}$. Convexity, expansion of the Bregman divergence, and (27) give

$$0 \le D_\phi\left(v - \frac{\nabla\phi(v) - \nabla\phi(u)}{C}, u\right) \le D_\phi(v, u) - \frac{1}{2C}\|\nabla\phi(v) - \nabla\phi(u)\|^2.$$

Interchanging u, v and adding yields

$$\|\nabla\phi(v) - \nabla\phi(u)\|^2 \le C\Delta_\phi(u, v) \le C\|\nabla\phi(v) - \nabla\phi(u)\|\|u - v\|.$$

Thus $\nabla\phi$ is C-Lipschitz continuous on $\mathcal{U}$.

By (25), ∇f is Lipschitz continuous on $\mathcal{V}$. Consequently, $\nabla\phi \circ \nabla f$ is Lipschitz continuous on $\mathcal{V} \cap (\nabla f)^{-1}(\mathcal{U})$, a neighborhood of $\bar{x}$ by continuity of ∇f. Since $\bar{x} \in \mathbb{R}^d$ was arbitrary, this composition is locally Lipschitz continuous on $\mathbb{R}^d$.

□

The two regularity conditions in Assumption 4(iii) yield the following common bounds.

Proposition 18. *Let $\eta \ge 0$ and $f \in C^2(\mathbb{R}^d)$ with $\nabla^2 g_\eta \succ 0$ on $\mathbb{R}^d$. Suppose that $\phi \in \mathcal{P}$ and Assumption 4(iii) holds. Then the following hold.*

(i) *g_η is locally strongly convex.*

(ii) *$\nabla\phi \circ \nabla f$ is locally Lipschitz on $\mathbb{R}^d$.*

(iii) *For every nonempty compact set $\mathcal{K} \subset \mathbb{R}^d$, there exists $L_\mathcal{K} > 0$ such that*

$$D_\phi(\nabla f(x), \nabla f(y)) \le L_\mathcal{K} D_{g_\eta}(y, x), \qquad x, y \in \mathcal{K}. \tag{28}$$

Consequently,

$$\Delta_\phi(\nabla f(x), \nabla f(y)) \le L_\mathcal{K} \Delta_{g_\eta}(x, y), \qquad x, y \in \mathcal{K}. \tag{29}$$

Proof of Proposition 18. Part (i) follows from Lemma 16 applied to g_η. If $\eta = 0$, part (ii) follows from Lemma 17, while (28) is the assumed local dual relative smoothness, since $g_0 = f$.

If $\eta > 0$, ∇f is locally Lipschitz because $f \in C^2$, and $\nabla\phi$ is locally Lipschitz on Ω_ϕ by Assumption 4(iii). Their composition is therefore locally Lipschitz (Rockafellar & Wets, 1997, Exercise 9.8(c)). Fix a nonempty compact $\mathcal{K} \subset \mathbb{R}^d$. Choose a Lipschitz constant L_f of ∇f on a compact convex set containing $\mathcal{K}$, and a Lipschitz constant L_ϕ of $\nabla\phi$ on a compact convex subset of Ω_ϕ containing $\nabla f(\mathcal{K})$ (Rockafellar & Wets, 1997, Theorem 9.2). Part (i) gives $m_g > 0$ such that

$\nabla^2 g_\eta \succeq m_g I$ on a compact convex set containing $\mathcal{K}$. The descent lemma (Beck, 2017, Lemma 5.7) yields

$$D_\phi(\nabla f(x), \nabla f(y)) \le \frac{L_\phi}{2}\|\nabla f(x) - \nabla f(y)\|^2 \le \frac{L_\phi L_f^2}{2}\|x - y\|^2 \le \frac{L_\phi L_f^2}{m_g} D_{g_\eta}(y, x).$$

This proves (28). In either case, adding it with x, y interchanged gives (29). □

For $\eta = 0$, (28) is local dual relative smoothness as defined in Definition 3. For $\eta > 0$, its left-hand side still uses ∇f, while the right-hand side uses the convexified function g_η; it is a consequence of the regularity assumptions rather than an additional hypothesis. The locally Lipschitz Hessian is used separately in the weakly convex stepsize analysis.

A.2 Proof of Lemma 5

Proof of Lemma 5. We adapt the argument of Ou et al. (2026, Lemma 3.1) to DPGD. Since (4) gives $T_k(x_{k-1}) = x_k$,

$$T_k(x_{k-1}) - T_k(x_k) = \gamma_k \nabla\phi(\nabla f(x_k)).$$

Expanding $D_\phi(\nabla f(x_{k-1}), \nabla f(x_k))$, we obtain

$$\begin{aligned} D_\phi(\nabla f(x_{k-1}), \nabla f(x_k)) &= \phi(\nabla f(x_{k-1})) - \phi(\nabla f(x_k)) \\ &\quad - \frac{1}{\gamma_k}\langle T_k(x_{k-1}) - T_k(x_k), \nabla f(x_{k-1}) - \nabla f(x_k)\rangle. \end{aligned}$$

By the definition of T_k, the last inner product is

$$\begin{aligned} &\langle T_k(x_{k-1}) - T_k(x_k), \nabla f(x_{k-1}) - \nabla f(x_k)\rangle \\ &\quad = \Delta_f(x_k, x_{k-1}) - \gamma_k \Delta_\phi(\nabla f(x_k), \nabla f(x_{k-1})). \end{aligned}$$

Combining these expressions and using G_k gives (9).

Next, suppose that $\nabla f(\bar{x}) = 0$. By (4),

$$0 = \langle \nabla\phi(\nabla f(x_k)), \nabla f(\bar{x}) - \nabla f(x_{k+1})\rangle - \frac{1}{\gamma_{k+1}}\langle x_k - x_{k+1}, \nabla f(\bar{x}) - \nabla f(x_{k+1})\rangle.$$

For the first term, (8) and $\gamma_{k+1} = \beta_{k+1}\gamma_k$ give

$$\langle \nabla\phi(\nabla f(x_k)), \nabla f(\bar{x}) - \nabla f(x_{k+1})\rangle = -G_k - D_\phi(0, \nabla f(x_k)) + \frac{1}{\gamma_{k+1}}\chi_{k+1}.$$

The algebraic three-point identity (2) also gives

$$-\langle x_k - x_{k+1}, \nabla f(\bar{x}) - \nabla f(x_{k+1})\rangle = D_f(x_k, \bar{x}) - D_f(x_{k+1}, \bar{x}) - D_f(x_k, x_{k+1}).$$

Substituting these expressions and rearranging yields (10). □

B Proofs for the Convex Analysis

Throughout this appendix, $\eta = 0$ and the algorithm uses (7). We omit the superscript 0 from the local quantities, as in Section 4.1.

B.1 Proof of the convex Lyapunov inequality

Write $\beta_{\max} := (1 + \sqrt{5})/2$. Since $\beta_0 = 1$, induction using the update rule gives

$$0 < \beta_k \le \bar{\beta}_k, \qquad 1 < \bar{\beta}_k \le \beta_{\max}, \qquad k \ge 1. \tag{30}$$

The definition of ζ_k gives

$$\Delta_f(x_k, x_{k-1}) = \frac{1 + \zeta_k}{\zeta_k} D_f(x_{k-1}, x_k).$$

Proof of Lemma 6. We adapt the argument of Ou et al. (2026, Lemmas 3.2 and 3.6) to DPGD. Multiply (9) in Lemma 5 by $\gamma_{k+1}\bar{\beta}_{k+1}$ and apply (10) with $\bar{x} = x_\star$. Convexity of ϕ gives

$$D_\phi(0, \nabla f(x_k)) \geq 0, \qquad D_\phi(\nabla f(x_{k-1}), \nabla f(x_k)) \geq 0,$$

so

$$\begin{aligned} &D_f(x_{k+1}, x_\star) + \gamma_{k+1}(1+\bar{\beta}_{k+1})G_k + D_f(x_k, x_{k+1}) \\ &\quad \leq D_f(x_k, x_\star) + \gamma_{k+1}\bar{\beta}_{k+1}G_{k-1} - \beta_{k+1}\bar{\beta}_{k+1}(1-\gamma_k\widehat{L}_k^{\mathrm{d}})\Delta_f(x_k, x_{k-1}) + \chi_{k+1}. \end{aligned} \tag{31}$$

Applying (2) with $h = f$, $x = x_k$, $y = x_{k+1}$, and $z = x_k - 2\bar{\beta}_{k+1}[T_k(x_{k-1}) - T_k(x_k)]$, and using convexity of f, gives

$$\chi_{k+1} \leq \frac{\beta_{k+1}}{2\bar{\beta}_{k+1}} D_f(x_k, x_{k+1}) + \beta_{k+1}\bar{\beta}_{k+1}\mathcal{R}_{k,2\bar{\beta}_{k+1}}\Delta_f(x_k, x_{k-1}). \tag{32}$$

Substituting (32) into (31) yields

$$\begin{aligned} &D_f(x_{k+1}, x_\star) + \gamma_{k+1}(1+\bar{\beta}_{k+1})G_k + \left(1 - \frac{\beta_{k+1}}{2\bar{\beta}_{k+1}}\right) D_f(x_k, x_{k+1}) \\ &\quad \leq D_f(x_k, x_\star) + \gamma_{k+1}\bar{\beta}_{k+1}G_{k-1} \\ &\qquad + \beta_{k+1}\bar{\beta}_{k+1}\left(\mathcal{R}_{k,2\bar{\beta}_{k+1}} - (1-\gamma_k\widehat{L}_k^{\mathrm{d}})\right)\Delta_f(x_k, x_{k-1}). \end{aligned} \tag{33}$$

Meanwhile, (7) ensures

$$\beta_{k+1}\bar{\beta}_{k+1}\left[\mathcal{R}_{k,2\bar{\beta}_{k+1}} - (1-\gamma_k\widehat{L}_k^{\mathrm{d}})\right]_+ \frac{1+\zeta_k}{\zeta_k} \leq \frac{1}{2}. \tag{34}$$

Using $\Delta_f(x_k, x_{k-1}) = \frac{1+\zeta_k}{\zeta_k} D_f(x_{k-1}, x_k)$, $a \leq [a]_+$, and $\beta_k \leq \bar{\beta}_k$, (34) gives

$$\begin{aligned} &\beta_{k+1}\bar{\beta}_{k+1}\left(\mathcal{R}_{k,2\bar{\beta}_{k+1}} - (1-\gamma_k\widehat{L}_k^{\mathrm{d}})\right)\Delta_f(x_k, x_{k-1}) \\ &\quad \leq \beta_{k+1}\bar{\beta}_{k+1}\left[\mathcal{R}_{k,2\bar{\beta}_{k+1}} - (1-\gamma_k\widehat{L}_k^{\mathrm{d}})\right]_+ \frac{1+\zeta_k}{\zeta_k} D_f(x_{k-1}, x_k) \\ &\quad \leq \frac{1}{2} D_f(x_{k-1}, x_k) \leq \left(1 - \frac{\beta_k}{2\bar{\beta}_k}\right) D_f(x_{k-1}, x_k). \end{aligned} \tag{35}$$

Combining (33) and (35), and using $\gamma_{k+1} = \beta_{k+1}\gamma_k$, yields

$$\mathcal{E}_{k+1} \leq \mathcal{E}_k - \gamma_k\big(1 + \bar{\beta}_k - \beta_{k+1}\bar{\beta}_{k+1}\big)G_{k-1}.$$

Finally, $\beta_{k+1} \leq \bar{\beta}_{k+1}$ and $\bar{\beta}_{k+1}^2 = 1 + \beta_k$ imply

$$1 + \bar{\beta}_k - \beta_{k+1}\bar{\beta}_{k+1} \geq 1 + \bar{\beta}_k - \bar{\beta}_{k+1}^2 = \bar{\beta}_k - \beta_k \geq 0.$$

□

B.2 Consequences of the Lyapunov inequality

We establish the technical lemmas needed for the convergence guarantees. For $K \geq 0$, write

$$G_K^{\min} := \min_{0 \leq k \leq K} G_k.$$

First, Lemma 6 implies the following.

Lemma 19. *Under the update rule in Algorithm 1 and Assumption 4, the following hold.*

(i) $(\mathcal{E}_k)_{k \geq 1}$ *is nonincreasing and bounded below, hence convergent.*

(ii) *For every* $K \geq 1$, $G_K^{\min} \leq \frac{\mathcal{E}_1}{\sum_{k=1}^{K+1}\gamma_k}$.

Proof. We adapt the argument of Ou et al. (2026, Lemma 4.1) to DPGD. Since $G_k \geq 0$, we have $\mathcal{E}_k \geq 0$, and the first claim follows from Lemma 6. In particular,

$$0 \leq \mathcal{E}_{k+1} \leq \mathcal{E}_k - \gamma_k(1 + \bar{\beta}_k - \bar{\beta}_{k+1}\beta_{k+1})G_{k-1}.$$

Summing from $k = 1$ to K gives

$$G_K^{\min} \sum_{k=1}^{K} \gamma_k\big(1 + \bar{\beta}_k - \beta_{k+1}\bar{\beta}_{k+1}\big) \leq \sum_{k=1}^{K} \gamma_k\big(1 + \bar{\beta}_k - \beta_{k+1}\bar{\beta}_{k+1}\big)G_{k-1} \leq \mathcal{E}_1 - \mathcal{E}_{K+1}. \tag{36}$$

The definition of $\mathcal{E}_{K+1}$ and nonnegativity of its terms give

$$\mathcal{E}_{K+1} \geq \gamma_{K+1}\big(1 + \bar{\beta}_{K+1}\big)G_K \geq \gamma_{K+1}\big(1 + \bar{\beta}_{K+1}\big)G_K^{\min}. \tag{37}$$

Combining (36) and (37) yields

$$\mathcal{E}_1 \geq G_K^{\min}\left[\sum_{k=1}^{K} \gamma_k\big(1 + \bar{\beta}_k - \beta_{k+1}\bar{\beta}_{k+1}\big) + \gamma_{K+1}\big(1 + \bar{\beta}_{K+1}\big)\right]. \tag{38}$$

Since $\gamma_{k+1} = \beta_{k+1}\gamma_k$,

$$\begin{aligned}
&\sum_{k=1}^{K} \gamma_k\big(1 + \bar{\beta}_k - \beta_{k+1}\bar{\beta}_{k+1}\big) + \gamma_{K+1}\big(1 + \bar{\beta}_{K+1}\big) \\
&\quad = \sum_{k=1}^{K}\big(\gamma_k + \gamma_k\bar{\beta}_k - \gamma_{k+1}\bar{\beta}_{k+1}\big) + \gamma_{K+1} + \gamma_{K+1}\bar{\beta}_{K+1} = \gamma_1\bar{\beta}_1 + \sum_{k=1}^{K+1} \gamma_k \geq \sum_{k=1}^{K+1} \gamma_k.
\end{aligned}$$

Thus (38) gives

$$G_K^{\min} \leq \frac{\mathcal{E}_1}{\gamma_1\bar{\beta}_1 + \sum_{k=1}^{K+1} \gamma_k} \leq \frac{\mathcal{E}_1}{\sum_{k=1}^{K+1} \gamma_k},$$

proving the second claim. □

This result reduces the convergence argument to showing $\sum_{k=1}^{\infty} \gamma_k = \infty$.

B.3 Local estimates for vanishing stepsizes

The next lemma gives the limit of the curvature ratio for the forward map along subsequences whose stepsizes approach zero.

Lemma 20. *Under Assumption 4, suppose that the sequence $\{x_k\}$ generated by Algorithm 1 lies in a compact set $\mathcal{C} \subset \mathbb{R}^d$. Then, for any index sequence $\{k_j\}$,*

$$\gamma_{k_j} \to 0 \tag{39}$$

implies

$$\mathcal{R}_{k_j, 2\bar{\beta}_{k_j+1}} \to 1. \tag{40}$$

Proof. We adapt the argument of Ou et al. (2026, Lemma 4.2) to DPGD. Set $u_k := x_k - x_{k-1}$ and $v_k := \nabla\phi(\nabla f(x_k)) - \nabla\phi(\nabla f(x_{k-1}))$. Then $T_k(x_k) - T_k(x_{k-1}) = u_k - \gamma_k v_k$. By Lemma 17, $\nabla\phi \circ \nabla f$ is Lipschitz continuous and bounded on $\mathcal{C}$. Hence there exist $L_F, M_F > 0$ such that

$$\|v_k\| \leq L_F\|u_k\|, \qquad \|u_k\| = \gamma_k\|\nabla\phi(\nabla f(x_{k-1}))\| \leq M_F\gamma_k. \tag{41}$$

By (39), we may assume $\gamma_{k_j}L_F \leq 1$ for sufficiently large j. Thus (41) gives

$$\|T_{k_j}(x_{k_j}) - T_{k_j}(x_{k_j-1})\| \leq \|u_{k_j}\| + \gamma_{k_j}\|v_{k_j}\| \leq 2M_F\gamma_{k_j} \to 0. \tag{42}$$

Moreover, (30) gives $2\bar{\beta}_{k_j+1} \leq 2\beta_{\max}$, and therefore

$$2\bar{\beta}_{k_j+1}\|T_{k_j}(x_{k_j}) - T_{k_j}(x_{k_j-1})\| \to 0.$$

Thus we can choose ξ_j on the segment between x_{k_j} and $x_{k_j} + 2\bar{\beta}_{k_j+1}(T_{k_j}(x_{k_j}) - T_{k_j}(x_{k_j-1}))$, and $\widetilde{\xi}_j$ on the segment between x_{k_j} and x_{k_j-1}, such that

$$\begin{aligned}\mathcal{R}_{k_j,2\bar{\beta}_{k_j+1}} &= \frac{\langle \nabla^2 f(\xi_j)(u_{k_j} - \gamma_{k_j} v_{k_j}), u_{k_j} - \gamma_{k_j} v_{k_j}\rangle}{\langle \nabla^2 f(\widetilde{\xi}_j) u_{k_j}, u_{k_j}\rangle} \\ &= \frac{\langle \nabla^2 f(\xi_j) u_{k_j}, u_{k_j}\rangle}{\langle \nabla^2 f(\widetilde{\xi}_j) u_{k_j}, u_{k_j}\rangle} - 2\gamma_{k_j} \frac{\langle \nabla^2 f(\xi_j) u_{k_j}, v_{k_j}\rangle}{\langle \nabla^2 f(\widetilde{\xi}_j) u_{k_j}, u_{k_j}\rangle} + \gamma_{k_j}^2 \frac{\langle \nabla^2 f(\xi_j) v_{k_j}, v_{k_j}\rangle}{\langle \nabla^2 f(\widetilde{\xi}_j) u_{k_j}, u_{k_j}\rangle}. \end{aligned} \tag{43}$$

Equations (41) and (42) give $u_{k_j} \to 0$ and

$$\|\xi_j - \widetilde{\xi}_j\| \le 2\bar{\beta}_{k_j+1}\|T_{k_j}(x_{k_j}) - T_{k_j}(x_{k_j-1})\| + \|u_{k_j}\| \to 0.$$

Hence a compact convex set $\mathcal{K}$ contains $\mathcal{C}, \{\xi_j\}, \{\widetilde{\xi}_j\}$. Apply Lemma 16 with $h = f$ and denote the Hessian bounds on $\mathcal{K}$ by $m_{\mathcal{K}}^f, M_{\mathcal{K}}^f > 0$. Equations (43) and (41) give

$$\left|\mathcal{R}_{k_j,2\bar{\beta}_{k_j+1}} - 1\right| \le \frac{1}{m_{\mathcal{K}}^f}\|\nabla^2 f(\xi_j) - \nabla^2 f(\widetilde{\xi}_j)\| + 2\gamma_{k_j}\frac{M_{\mathcal{K}}^f L_F}{m_{\mathcal{K}}^f} + \gamma_{k_j}^2 \frac{M_{\mathcal{K}}^f L_F^2}{m_{\mathcal{K}}^f}. \tag{44}$$

Since $\nabla^2 f$ is uniformly continuous on $\mathcal{K}$ and $\|\xi_j - \widetilde{\xi}_j\| \to 0$, the right-hand side of (44) tends to 0. This proves (40). □

B.4 Iterate boundedness and nonsummability of stepsizes

Proposition 21. *Under the update rule in Algorithm 1 and Assumption 4, the following hold.*

(i) *There is a compact set $\mathcal{K}$ containing $\{x_k\} \subset \mathbb{R}^d$. Consequently, a compact set $\mathcal{K}^* \subset \Omega_\phi$ also contains $\{\nabla f(x_k)\}$.*

(ii)
$$\sum_{k=1}^{\infty} \gamma_k = \infty. \tag{45}$$

(iii) *A subsequence of $\{x_k\}$ converges to $x_\star$. Correspondingly, a subsequence of $\{\nabla f(x_k)\}$ converges to 0.*

Proof. (i) By Lemma 19, for all $k \ge 1$,

$$D_f(x_k, x_\star) \le \mathcal{E}_k \le \mathcal{E}_1.$$

Hence

$$x_k \in \{x \in \mathbb{R}^d | f(x) \le f(x_\star) + \mathcal{E}_1\}, \qquad k \ge 1.$$

The right-hand side is compact because f is level bounded, and adjoining x_0 preserves compactness. This proves the claim; the statement for the dual sequence follows by continuity of ∇f.

(ii) Suppose, toward a contradiction, that (45) fails. Then $\sum_k \gamma_k < \infty$, so $\gamma_k \to 0$. Choose a compact convex set $\mathcal{K}$ containing the compact set in Proposition 21(i). By (24) in Lemma 16, $\zeta_k \ge \bar{\zeta}$ for some $\bar{\zeta} > 0$. Applying local dual relative smoothness in both orders to x_k, x_{k-1} gives $\widehat{L}_k^{\mathrm{d}} \le \overline{L}^{\mathrm{d}}$ for some $\overline{L}^{\mathrm{d}} > 0$. Applying Lemma 20 with $k_j = j$ yields

$$\mathcal{R}_{k,2\bar{\beta}_{k+1}} \to 1.$$

The second candidate in (7) therefore satisfies

$$\frac{\zeta_k}{1+\zeta_k}\frac{1}{2\bar{\beta}_{k+1}[\mathcal{R}_{k,2\bar{\beta}_{k+1}} - (1-\gamma_k \widehat{L}_k^{\mathrm{d}})]_+} \ge \frac{\bar{\zeta}}{1+\bar{\zeta}}\frac{1}{2\beta_{\max}[\mathcal{R}_{k,2\bar{\beta}_{k+1}} - (1-\gamma_k \overline{L}^{\mathrm{d}})]_+} \to \infty,$$

where $[\mathcal{R}_{k,2\bar{\beta}_{k+1}} - (1-\gamma_k \overline{L}^{\mathrm{d}})]_+ \to 0$ follows from $\gamma_k \to 0$ and the preceding limit. Thus $\beta_{k+1} = \bar{\beta}_{k+1}$ for sufficiently large k. Advancing the index once more gives $\beta_k = \bar{\beta}_k > 1$, so

$$\bar{\beta}_{k+1} = \sqrt{1+\beta_k} \ge \sqrt{2}$$

for all sufficiently large k. Hence $\gamma_{k+1} = \bar{\beta}_{k+1}\gamma_k \geq \sqrt{2}\gamma_k$, forcing γ_k to diverge, contrary to $\gamma_k \to 0$.

(iii) Proposition 21(ii) and Lemma 19 give $G_K^{\min} \to 0$. Choose indices with $G_{k_j} \to 0$. By Proposition 21(i), pass to a further subsequence if necessary so that $x_{k_j} \to \bar{x}$. Continuity of $\phi \circ \nabla f$ and Assumption 4 give $\bar{x} = x_\star$, and hence $\nabla f(x_{k_j}) \to 0$. □

In view of Proposition 21(iii), it remains to prove $x_k \to x_\star$. We now use these auxiliary results to give complete proofs of the main results.

B.5 Convergence of the entire sequence

Proof of Theorem 7. We adapt the argument of Ou et al. (2026, Theorem 4.4) to DPGD. By Lemma 19, $\{\mathcal{E}_k\}$ is nonnegative and nonincreasing. Thus there exists $\mathcal{E} \geq 0$ such that

$$\mathcal{E}_k \to \mathcal{E}. \tag{46}$$

We show that $\mathcal{E} = 0$.

By Proposition 21(iii), there exists a subsequence $\{x_{k_j}\}$ such that

$$x_{k_j} \to x_\star. \tag{47}$$

We distinguish cases according to the corresponding stepsizes $\{\gamma_{k_j+1}\}$.

First suppose that $\{\gamma_{k_j+1}\}$ has a bounded subsequence. Pass to that subsequence and retain the notation. Since $\nabla\phi(\nabla f(x_\star)) = 0$, continuity of $\nabla\phi(\nabla f(\cdot))$ and the update give

$$\|x_{k_j+1} - x_\star\| \leq \|x_{k_j} - x_\star\| + \gamma_{k_j+1}\|\nabla\phi(\nabla f(x_{k_j}))\| \to 0.$$

By (47) and continuity of $\phi \circ \nabla f$,

$$G_{k_j} \to 0.$$

Since $\{\bar{\beta}_{k_j+1}\}$ is also bounded,

$$\gamma_{k_j+1}\big(1 + \bar{\beta}_{k_j+1}\big)G_{k_j} \to 0.$$

Furthermore,

$$D_f(x_{k_j+1}, x_\star) \to 0, \qquad D_f(x_{k_j}, x_{k_j+1}) \to 0,$$

so the definition of the Lyapunov function gives

$$\mathcal{E}_{k_j+1} \to 0.$$

Equation (46) therefore implies $\mathcal{E} = 0$.

Next suppose that $\{\gamma_{k_j+1}\}$ has no bounded subsequence. Then

$$\gamma_{k_j+1} \to \infty. \tag{48}$$

The upper bound on the stepsize ratio,

$$\gamma_{k+1} \leq \beta_{\max}\gamma_k,$$

together with (48), gives

$$\gamma_{k_j} \geq \frac{\gamma_{k_j+1}}{\beta_{\max}} \to \infty. \tag{49}$$

Nonnegativity and boundedness of the Lyapunov function give

$$0 \leq \gamma_{k_j}G_{k_j-1} \leq \mathcal{E}_{k_j} \leq \mathcal{E}_1.$$

Hence (49) yields

$$G_{k_j-1} \to 0. \tag{50}$$

By Proposition 21(i), $\{x_{k_j-1}\}$ is bounded. For any cluster point $\bar{x}$, (50) and continuity of $\phi \circ \nabla f$ give

$$\phi(\nabla f(\bar{x})) = \phi(0).$$

By Assumption 4, $\bar{x} = x_\star$. Thus

$$x_{k_j-1} \to x_\star. \tag{51}$$

Convexity of ϕ gives, for any $y \in \Omega_\phi$,

$$\phi(y) - \phi(0) \le \langle \nabla\phi(y), y\rangle.$$

Applying this to $y = \nabla f(x_{k_j-1})$ and using the DPGD update yields

$$\gamma_{k_j} G_{k_j-1} \le \gamma_{k_j} \langle \nabla\phi(\nabla f(x_{k_j-1})), \nabla f(x_{k_j-1})\rangle = \langle x_{k_j-1} - x_{k_j}, \nabla f(x_{k_j-1})\rangle \to 0. \tag{52}$$

The last limit uses (47), (51), and $\nabla f(x_\star) = 0$.

Since $\{\bar{\beta}_k\}$ is bounded and

$$0 \le 1 - \frac{\beta_k}{2\bar{\beta}_k} \le 1,$$

(47), (51), (52), and the definition of the Lyapunov function give

$$\mathcal{E}_{k_j} = D_f(x_{k_j}, x_\star) + \gamma_{k_j}\big(1 + \bar{\beta}_{k_j}\big) G_{k_j-1} + \left(1 - \frac{\beta_{k_j}}{2\bar{\beta}_{k_j}}\right) D_f(x_{k_j-1}, x_{k_j}) \to 0.$$

Thus $\mathcal{E} = 0$ in this case as well.

In either case,

$$\mathcal{E}_k \to 0.$$

In particular,

$$0 \le f(x_k) - f(x_\star) = D_f(x_k, x_\star) \le \mathcal{E}_k \to 0.$$

Boundedness from Proposition 21(i), continuity of f, and uniqueness of $x_\star$ imply

$$x_k \to x_\star.$$

Continuity of ∇f then gives $\nabla f(x_k) \to 0$.

□

B.6 Linear convergence

Proof of Theorem 9. If a stationary point is reached, it is $x_\star$ by Assumption 4; extending the iterates constantly gives (12) after enlarging the constants to cover the finite prefix. We therefore consider an infinite nonstationary sequence.

By Proposition 21(i), choose a compact convex set $\mathcal{C} \subset \mathbb{R}^d$ containing $x_\star$ and every iterate. The set $\nabla f(\mathcal{C})$ is a compact subset of the open convex set Ω_ϕ, so choose a compact convex set $\mathcal{K}^* \subset \Omega_\phi$ containing it. In particular, $0 = \nabla f(x_\star) \in \mathcal{K}^*$. Write $\mu_\phi := \mu_{\mathcal{K}^*} > 0$ from Assumption 8. By Lemmas 16 and 17, there are $m_f, M_f, L_F > 0$ such that

$$m_f I \preceq \nabla^2 f(x) \preceq M_f I, \qquad \|\nabla\phi(\nabla f(x)) - \nabla\phi(\nabla f(y))\| \le L_F \|x - y\|, \qquad x, y \in \mathcal{C}, \tag{53}$$

With $\beta_{\max} = (1 + \sqrt{5})/2$, initialization and (30) give

$$0 < \beta_k \le \bar{\beta}_k \le \beta_{\max}, \qquad \bar{\beta}_k > 1, \qquad \bar{\beta}_{k+1}^2 = 1 + \beta_k, \qquad k \ge 1. \tag{54}$$

Strong convexity of ϕ on $\mathcal{K}^*$ and the local Hessian bounds yield

$$\Delta_\phi(\nabla f(x_k), \nabla f(x_{k-1})) \ge \mu_\phi \|\nabla f(x_k) - \nabla f(x_{k-1})\|^2,$$
$$\Delta_f(x_k, x_{k-1}) \le \|\nabla f(x_k) - \nabla f(x_{k-1})\| \|x_k - x_{k-1}\| \le m_f^{-1} \|\nabla f(x_k) - \nabla f(x_{k-1})\|^2.$$

Applying local dual relative smoothness in both orders and the quadratic Bregman bounds on $\mathcal{C}$ therefore gives a constant $\overline{L}^{\mathrm{d}} > 0$ such that

$$0 < \underline{L}^{\mathrm{d}} := \mu_\phi m_f \le \widehat{L}_k^{\mathrm{d}} \le \overline{L}^{\mathrm{d}}, \qquad \zeta_k \ge \underline{\zeta} := m_f/M_f > 0. \tag{55}$$

We first bound the stepsizes. If $\gamma_k \ge 2/\underline{L}^{\mathrm{d}}$, nonnegativity of $\mathcal{R}_{k,2\bar{\beta}_{k+1}}$ gives

$$\mathcal{R}_{k,2\bar{\beta}_{k+1}} - (1 - \gamma_k \widehat{L}_k^{\mathrm{d}}) \ge \gamma_k \underline{L}^{\mathrm{d}} - 1 \ge \tfrac{1}{2}\gamma_k \underline{L}^{\mathrm{d}}.$$

Thus (7) implies $\gamma_{k+1} \le 1/\underline{L}^{\mathrm{d}}$. Otherwise, $\gamma_{k+1} \le \beta_{\max}\gamma_k < 2\beta_{\max}/\underline{L}^{\mathrm{d}}$. Since $\gamma_1 = \gamma_0$,

$$\gamma_k \le \overline{\gamma} := \max\{\gamma_0, 2\beta_{\max}/\underline{L}^{\mathrm{d}}\}, \qquad k \ge 0. \tag{56}$$

Because $\nabla\phi \circ \nabla f$ is bounded on $\mathcal{C}$, the points $z_k := x_k - 2\bar{\beta}_{k+1}\gamma_k \nabla\phi(\nabla f(x_k))$ are bounded. Choose a compact convex set $\mathcal{Q}$ containing $\mathcal{C}$ and all z_k, and let $M_{\mathcal{Q}}^f > 0$ bound the largest eigenvalue of $\nabla^2 f$ on $\mathcal{Q}$. The update and (53) imply

$$\|\nabla\phi(\nabla f(x_k))\| \le \|\nabla\phi(\nabla f(x_{k-1}))\| + L_F\|x_k - x_{k-1}\| \le (1 + L_F\overline{\gamma})\|\nabla\phi(\nabla f(x_{k-1}))\|.$$

Since $T_k(x_{k-1}) = x_k$, the point z_k is the evaluation point in (5) for $\delta = 2\bar{\beta}_{k+1}$. Using $x_k - x_{k-1} = -\gamma_k\nabla\phi(\nabla f(x_{k-1}))$ and $\nabla\phi(\nabla f(x_{k-1})) \ne 0$, the quadratic bounds give

$$\mathcal{R}_{k,2\bar{\beta}_{k+1}} = \frac{2D_f(z_k, x_k)}{4\bar{\beta}_{k+1}^2 \Delta_f(x_k, x_{k-1})} \le \frac{M_{\mathcal{Q}}^f}{m_f}\frac{\|\nabla\phi(\nabla f(x_k))\|^2}{\|\nabla\phi(\nabla f(x_{k-1}))\|^2} \le \frac{M_{\mathcal{Q}}^f}{m_f}(1 + L_F\overline{\gamma})^2 =: \overline{\mathcal{R}}. \tag{57}$$

Combining (55)–(57) with the stepsize rule, including its division convention, yields

$$\beta_{k+1} \ge \underline{\beta} := \min\left\{1, \frac{\underline{\zeta}}{1+\underline{\zeta}}\frac{1}{2\beta_{\max}(\overline{\mathcal{R}} + \overline{\gamma}\,\overline{L}^{\mathrm{d}})}\right\} > 0. \tag{58}$$

Moreover, there exists $\gamma_{\mathrm{safe}} > 0$ such that

$$0 < \gamma_k \le \gamma_{\mathrm{safe}} \quad \Longrightarrow \quad \beta_{k+1} = \bar{\beta}_{k+1}. \tag{59}$$

Otherwise, there are indices k_j with $\gamma_{k_j} \to 0$ and $\beta_{k_j+1} < \bar{\beta}_{k_j+1}$. By Lemma 20 and (55),

$$[\mathcal{R}_{k_j,2\bar{\beta}_{k_j+1}} - (1 - \gamma_{k_j}\widehat{L}_{k_j}^{\mathrm{d}})]_+ \to 0.$$

Since $\zeta_{k_j} \ge \underline{\zeta} > 0$ and $\bar{\beta}_{k_j+1} \le \beta_{\max}$, the second candidate in (7) tends to $+\infty$, contradicting the strict inequality. Set $\underline{\gamma} := \min\{\gamma_1, \underline{\beta}\gamma_{\mathrm{safe}}\} > 0$. If $\gamma_k \le \gamma_{\mathrm{safe}}$, (59) gives $\gamma_{k+1} \ge \gamma_k$; otherwise, (58) gives $\gamma_{k+1} \ge \underline{\beta}\gamma_{\mathrm{safe}}$. Induction and (56) therefore yield

$$0 < \underline{\gamma} \le \gamma_k \le \overline{\gamma} < \infty, \qquad k \ge 1.$$

Next we compare the Lyapunov energy with G_{k-1}. Since 0 minimizes ϕ and belongs to Ω_ϕ, $\nabla\phi(0) = 0$ and $\nabla\phi(\nabla f(x_\star)) = 0$. Strong convexity on $\mathcal{K}^*$ and (53) give

$$G_j \ge \frac{\mu_\phi}{2}\|\nabla f(x_j)\|^2 \ge \frac{\mu_\phi m_f^2}{2}\|x_j - x_\star\|^2, \qquad j \ge 0. \tag{60}$$

The update gives

$$\|x_k - x_\star\| \le (1 + \overline{\gamma}L_F)\|x_{k-1} - x_\star\|, \qquad \|x_k - x_{k-1}\| \le \overline{\gamma}L_F\|x_{k-1} - x_\star\|.$$

Using $D_f(x_k, x_\star) = f(x_k) - f_\star$, $1 - \beta_k/(2\bar{\beta}_k) \in [1/2, 1]$, and the upper quadratic bounds in (11), we obtain

$$0 \le f(x_k) - f_\star \le \mathcal{E}_k \le \frac{M_f}{2}\left[(1+\overline{\gamma}L_F)^2 + \overline{\gamma}^2 L_F^2\right]\|x_{k-1} - x_\star\|^2 + \overline{\gamma}(1+\beta_{\max})G_{k-1} \le C_E G_{k-1}, \tag{61}$$

where the finite constant

$$C_E := \overline{\gamma}(1+\beta_{\max}) + \frac{M_f}{\mu_\phi m_f^2}\left[(1+\overline{\gamma}L_F)^2 + \overline{\gamma}^2 L_F^2\right] > 0$$

is independent of k.

Finally, set $c_k := 1 + \bar{\beta}_k - \beta_{k+1}\bar{\beta}_{k+1}$. By (54), $c_k \ge \bar{\beta}_k - \beta_k \ge 0$, and Lemma 6 gives

$$\mathcal{E}_{k+1} \le \mathcal{E}_k - \gamma_k c_k G_{k-1}. \tag{62}$$

For any $n \ge 1$ and integer $N \ge 1$, the identity $\gamma_k \beta_{k+1} = \gamma_{k+1}$ gives

$$\sum_{k=n}^{n+N-1} \gamma_k c_k = \sum_{k=n}^{n+N-1} \gamma_k + \gamma_n \bar{\beta}_n - \gamma_{n+N}\bar{\beta}_{n+N} \ge N\underline{\gamma} - \beta_{\max}\overline{\gamma}.$$

Fix N large enough that $N\underline{\gamma} > \beta_{\max}\overline{\gamma}$. Summing (62) and using (61), $\gamma_k c_k \ge 0$, and monotonicity of the energy yields

$$\mathcal{E}_n - \mathcal{E}_{n+N} \ge \sum_{k=n}^{n+N-1} \gamma_k c_k G_{k-1} \ge C_E^{-1} \sum_{k=n}^{n+N-1} \gamma_k c_k \mathcal{E}_k \ge \frac{N\underline{\gamma} - \beta_{\max}\overline{\gamma}}{C_E}\mathcal{E}_{n+N}.$$

Thus, with $\kappa := \left(1 + (N\underline{\gamma} - \beta_{\max}\overline{\gamma})/C_E\right)^{-1} \in (0,1)$ independent of n,

$$\mathcal{E}_{n+N} \le \kappa \mathcal{E}_n, \qquad \mathcal{E}_k \le \mathcal{E}_1 \kappa^{\lfloor (k-1)/N \rfloor}, \qquad n, k \ge 1. \tag{63}$$

Put $\alpha := \kappa^{1/(2N)} \in (0,1)$. Since $f(x_k) - f_\star \le \mathcal{E}_k$ and $\mathcal{E}_{k+1} \ge \underline{\gamma} G_k$, (63) and (60) imply

$$f(x_k) - f_\star \le \frac{\mathcal{E}_1}{\kappa\alpha^2}\alpha^{2k} \quad (k \ge 1), \qquad \|\nabla f(x_k)\|^2 \le \frac{2\mathcal{E}_1}{\kappa\underline{\gamma}\mu_\phi}\alpha^{2k} \quad (k \ge 0).$$

Enlarging C_f to cover $k = 0$ proves (12), since $\alpha^{2k} \le \alpha^k$. □

C Proofs for the Weakly Convex Analysis

Throughout this appendix, $\eta > 0$ and the algorithm uses (7).

C.1 Proof of the weakly convex Lyapunov inequality

Induction using the weakly convex stepsize rule gives

$$0 < \beta_k \le \bar{\beta}_k \le \beta_{\mathrm{cap}}. \tag{64}$$

The definitions of m_k^η and ζ_k^η give

$$\eta\|x_k - x_{k-1}\|^2 = \frac{\eta}{m_k^\eta}\Delta_{g_\eta}(x_k, x_{k-1}). \tag{65}$$

They also give

$$\Delta_{g_\eta}(x_k, x_{k-1}) = \frac{1+\zeta_k^\eta}{\zeta_k^\eta} D_{g_\eta}(x_{k-1}, x_k). \tag{66}$$

Proof of Lemma 10. Multiply (9) in Lemma 5 by $\gamma_{k+1}/2$ and apply (10) with $\bar{x} = x_\star$. Using (14) and $D_\phi \geq 0$ gives

$$\begin{aligned} &D_f(x_{k+1}, x_\star) + \frac{3}{2}\gamma_{k+1} G_k + D_f(x_k, x_{k+1}) \\ &\quad \leq D_f(x_k, x_\star) + \frac{1}{2}\gamma_{k+1} G_{k-1} - \frac{\beta_{k+1}}{2}\left(1 - \gamma_k \widehat{L}_k^{\mathrm{d},\eta} - \frac{\eta}{m_k^\eta}\right) \Delta_{g_\eta}(x_k, x_{k-1}) + \chi_{k+1}. \end{aligned} \tag{67}$$

By the update and the definition of χ_{k+1},

$$\chi_{k+1} = \langle \nabla g_\eta(x_{k+1}) - \nabla g_\eta(x_k), x_{k+1} - x_k \rangle - \eta \|x_{k+1} - x_k\|^2.$$

As in (32), convexity of g_η and (2) give

$$\chi_{k+1} \leq \frac{\beta_{k+1}}{\bar{\beta}_{k+1}} D_{g_\eta}(x_k, x_{k+1}) + \frac{\bar{\beta}_{k+1}\beta_{k+1}}{2} \mathcal{R}^\eta_{k,\bar{\beta}_{k+1}} \Delta_{g_\eta}(x_k, x_{k-1}) - \eta \|x_{k+1} - x_k\|^2. \tag{68}$$

Substituting (68) into (67) and using

$$D_f(x_k, x_{k+1}) = D_{g_\eta}(x_k, x_{k+1}) - \frac{\eta}{2}\|x_{k+1} - x_k\|^2$$

yields

$$\begin{aligned} &D_f(x_{k+1}, x_\star) + \frac{3}{2}\gamma_{k+1} G_k + \left(1 - \frac{\beta_{k+1}}{\bar{\beta}_{k+1}}\right) D_{g_\eta}(x_k, x_{k+1}) + \frac{\eta}{2}\|x_{k+1} - x_k\|^2 \\ &\quad \leq D_f(x_k, x_\star) + \frac{1}{2}\gamma_{k+1} G_{k-1} \\ &\qquad + \beta_{k+1}\left(\frac{\bar{\beta}_{k+1}}{2}\mathcal{R}^\eta_{k,\bar{\beta}_{k+1}} - \frac{1}{2}\left(1 - \gamma_k \widehat{L}_k^{\mathrm{d},\eta} - \frac{\eta}{m_k^\eta}\right)\right) \Delta_{g_\eta}(x_k, x_{k-1}). \end{aligned} \tag{69}$$

By (64), the numerator of the second candidate in (7) is nonnegative. Thus (7) gives

$$\begin{aligned} &\beta_{k+1}\left(\frac{\bar{\beta}_{k+1}}{2}\mathcal{R}^\eta_{k,\bar{\beta}_{k+1}} - \frac{1}{2}\left(1 - \gamma_k \widehat{L}_k^{\mathrm{d},\eta} - \frac{\eta}{m_k^\eta}\right)\right) \\ &\quad \leq \beta_{k+1}\left[\frac{\bar{\beta}_{k+1}}{2}\mathcal{R}^\eta_{k,\bar{\beta}_{k+1}} - \frac{1}{2}\left(1 - \gamma_k \widehat{L}_k^{\mathrm{d},\eta} - \frac{\eta}{m_k^\eta}\right)\right]_+ \\ &\quad \leq \frac{\zeta_k^\eta}{1+\zeta_k^\eta}\left(1 - \frac{\beta_k}{\bar{\beta}_k}\right) + \frac{\eta}{2m_k^\eta}. \end{aligned} \tag{70}$$

Multiply (70) by $\Delta_{g_\eta}(x_k, x_{k-1})$ and use (65) and (66) to obtain

$$\begin{aligned} &\beta_{k+1}\left(\frac{\bar{\beta}_{k+1}}{2}\mathcal{R}^\eta_{k,\bar{\beta}_{k+1}} - \frac{1}{2}\left(1 - \gamma_k \widehat{L}_k^{\mathrm{d},\eta} - \frac{\eta}{m_k^\eta}\right)\right) \Delta_{g_\eta}(x_k, x_{k-1}) \\ &\quad \leq \left(1 - \frac{\beta_k}{\bar{\beta}_k}\right) D_{g_\eta}(x_{k-1}, x_k) + \frac{\eta}{2}\|x_k - x_{k-1}\|^2. \end{aligned} \tag{71}$$

Substituting (71) into (69) and using $\gamma_{k+1} = \beta_{k+1}\gamma_k$ gives

$$\mathcal{E}_{k+1}^{\mathrm{weak}} \leq \mathcal{E}_k^{\mathrm{weak}} - \frac{\gamma_k}{2}(3 - \beta_{k+1}) G_{k-1},$$

which is (16). Finally, $\beta_{k+1} \leq \beta_{\mathrm{cap}} < 3$ yields

$$\mathcal{E}_{k+1}^{\mathrm{weak}} \leq \mathcal{E}_k^{\mathrm{weak}} - \frac{\gamma_k}{2}(3 - \beta_{\mathrm{cap}}) G_{k-1}. \tag{72}$$

□

C.2 Consequences of the Lyapunov inequality

By (13) and (64), every term in (15) is nonnegative. As in the convex analysis, the Lyapunov function first gives the following.

Lemma 22. *Under Assumption 4 with $\eta > 0$ and Algorithm 1, the following hold.*

(i) $\{\mathcal{E}_k^{\mathrm{weak}}\}$ *is nonnegative and nonincreasing, hence convergent.*

(ii) *For every $K \geq 1$,*

$$\min_{0\leq j\leq K-1} G_j \leq \frac{2}{3-\beta_{\mathrm{cap}}} \frac{\mathcal{E}_1^{\mathrm{weak}}}{\sum_{k=1}^K \gamma_k}. \tag{73}$$

Furthermore,

$$\sum_{k=1}^{\infty} \gamma_k G_{k-1} < \infty. \tag{74}$$

Proof. Equation (72) implies that $\mathcal{E}_k^{\mathrm{weak}}$ is nonnegative and nonincreasing, and for every $K \geq 1$,

$$\mathcal{E}_{K+1}^{\mathrm{weak}} + \frac{3-\beta_{\mathrm{cap}}}{2} \sum_{k=1}^{K} \gamma_k G_{k-1} \leq \mathcal{E}_1^{\mathrm{weak}}. \tag{75}$$

This proves the first claim; letting $K \to \infty$ gives (74). Moreover, (75) and $G_{k-1} \geq \min_{0\leq j\leq K-1} G_j$ yield

$$\frac{3-\beta_{\mathrm{cap}}}{2} \left(\min_{0\leq j\leq K-1} G_j \right) \sum_{k=1}^{K} \gamma_k \leq \mathcal{E}_1^{\mathrm{weak}},$$

proving (73). □

C.3 Local estimates for vanishing stepsizes

For convex objectives, Lemma 20 analyzes the limit of the local estimate along subsequences whose stepsizes tend to 0. The contradiction argument in Proposition 21(ii) then proves $\sum_k \gamma_k = \infty$. For weakly convex objectives, we do not rule out stepsizes approaching 0. Instead, we show that even if $\gamma_k \to 0$, the decrease is slow enough to preserve $\sum_k \gamma_k = \infty$. Besides local boundedness of $m_k^\eta, \widehat{L}_k^{\mathrm{d},\eta}$, the next lemma quantifies how $\mathcal{R}^\eta_{k,\bar{\beta}_{k+1}}$ approaches 1.

Lemma 23 (local estimates)**.** *Under Assumption 4 with $\eta > 0$, suppose that $\{x_k\}$ lies in a compact set and $\gamma_k \to 0$. Then there exist constants $0 < \underline{m} \leq \overline{m} < \infty$, $\overline{L}^{\mathrm{d}} < \infty$, $C_{\mathcal{R}} > 0$, and $k_0 \in \mathbb{N}$ such that, for all $k \geq k_0$,*

$$\underline{m} \leq m_k^\eta \leq \overline{m}, \qquad 0 \leq \widehat{L}_k^{\mathrm{d},\eta} \leq \overline{L}^{\mathrm{d}}$$

and

$$\left| \mathcal{R}^\eta_{k,\bar{\beta}_{k+1}} - 1 \right| \leq C_{\mathcal{R}} \gamma_k. \tag{76}$$

Proof. Let $\mathcal{C}$ be the compact convex hull of a compact set containing the iterates. Applying Lemma 16 to $h = g_\eta$ yields constants $\underline{m}, \overline{m} > 0$ such that, by (23),

$$\underline{m} \leq m_k^\eta = \frac{\Delta_{g_\eta}(x_k, x_{k-1})}{\|x_k - x_{k-1}\|^2} \leq \overline{m}.$$

Applying (29) on $\mathcal{C}$ gives $\overline{L}^{\mathrm{d}} > 0$ such that

$$0 \leq \widehat{L}_k^{\mathrm{d},\eta} \leq \overline{L}^{\mathrm{d}}.$$

It remains to prove (76). Set

$$u_k := x_k - x_{k-1}, \qquad v_k := \nabla\phi(\nabla f(x_k)) - \nabla\phi(\nabla f(x_{k-1})).$$

By Proposition 18(ii), $\nabla\phi \circ \nabla f$ is Lipschitz continuous and bounded on $\mathcal{C}$. Let $L_F > 0$ be its Lipschitz constant and $M_F > 0$ a bound on its norm on $\mathcal{C}$. The update gives

$$\|v_k\| \le L_F\|u_k\|, \qquad \|u_k\| = \gamma_k\|\nabla\phi(\nabla f(x_{k-1}))\| \le M_F\gamma_k. \tag{77}$$

The definition of $\bar\beta_{k+1}$ in the weakly convex setting also gives

$$1 \le \bar\beta_{k+1} \le \beta_{\mathrm{cap}}, \qquad 0 \le \bar\beta_{k+1} - 1 \le s\gamma_k. \tag{78}$$

Since $\gamma_k \to 0$, we may assume $\gamma_k L_F \le 1$ for sufficiently large k. Hence

$$\|T_k(x_k) - T_k(x_{k-1})\| = \|u_k - \gamma_k v_k\| \le 2M_F\gamma_k. \tag{79}$$

Setting $z_k := x_k + \bar\beta_{k+1}(T_k(x_k) - T_k(x_{k-1}))$, (78) and (79) give

$$\|z_k - x_k\| \le 2\beta_{\mathrm{cap}}M_F\gamma_k.$$

Thus there is a compact convex set $\mathcal{K}$ containing $\mathcal{C}$ and all z_k for sufficiently large k. Since g_η is C^2 with a locally Lipschitz continuous Hessian, Lemma 16 gives $m_g, M_g, L_g > 0$ such that, for all $x, y \in \mathcal{K}$,

$$m_g I \preceq \nabla^2 g_\eta(x) \preceq M_g I, \qquad \|\nabla^2 g_\eta(x) - \nabla^2 g_\eta(y)\| \le L_g\|x - y\|. \tag{80}$$

By Taylor's theorem, choose ξ_k on the segment joining x_k, z_k and $\widetilde{\xi}_k$ on the segment joining x_{k-1}, x_k such that

$$\mathcal{R}^\eta_{k,\bar\beta_{k+1}} = \frac{\langle\nabla^2 g_\eta(\xi_k)(u_k - \gamma_k v_k), u_k - \gamma_k v_k\rangle}{\langle\nabla^2 g_\eta(\widetilde{\xi}_k)u_k, u_k\rangle}.$$

Expanding the Hessian quotient for g_η as after (43), and using (77) and (80), gives

$$\left|\mathcal{R}^\eta_{k,\bar\beta_{k+1}} - 1\right| \le \frac{1}{m_g}\|\nabla^2 g_\eta(\xi_k) - \nabla^2 g_\eta(\widetilde{\xi}_k)\| + 2\gamma_k\frac{M_g L_F}{m_g} + \gamma_k^2\frac{M_g L_F^2}{m_g}. \tag{81}$$

The locations of $\xi_k, \widetilde{\xi}_k$ and (77), (78), and (79) imply

$$\|\xi_k - \widetilde{\xi}_k\| \le \bar\beta_{k+1}\|T_k(x_k) - T_k(x_{k-1})\| + \|u_k\| \le (2\beta_{\mathrm{cap}} + 1)M_F\gamma_k.$$

Hence (80) yields

$$\|\nabla^2 g_\eta(\xi_k) - \nabla^2 g_\eta(\widetilde{\xi}_k)\| \le L_g(2\beta_{\mathrm{cap}} + 1)M_F\gamma_k.$$

Increase k_0 so that $\gamma_k \le 1$. Equation (81) then gives

$$\left|\mathcal{R}^\eta_{k,\bar\beta_{k+1}} - 1\right| \le C_{\mathcal{R}}\gamma_k,$$

with

$$C_{\mathcal{R}} := \frac{L_g(2\beta_{\mathrm{cap}} + 1)M_F + 2M_g L_F + M_g L_F^2}{m_g}.$$

This proves the result. □

C.4 Iterate boundedness and nonsummability of stepsizes

We next prove the counterpart of Proposition 21 needed for the weakly convex analysis.

Proposition 24. *Under Assumption 4 with $\eta > 0$ and Algorithm 1, the following hold.*

(i) *There exists a compact set $\mathcal{C} \subset \mathbb{R}^d$ containing $\{x_k\}$.*

(ii)
$$\sum_{k=1}^{\infty}\gamma_k = \infty. \tag{82}$$

(iii)
$$\liminf_{k\to\infty} G_k = 0. \tag{83}$$

Proof. (i) Use the sublevel-set argument from Proposition 21(i). By Lemma 22,

$$f(x_k) - f(x_\star) = D_f(x_k, x_\star) \le \mathcal{E}_k^{\mathrm{weak}} \le \mathcal{E}_1^{\mathrm{weak}}, \qquad k \ge 1.$$

Hence $\{x_k\}_{k\ge1}$ lies in a single sublevel set of f. This set is compact by level boundedness in Assumption 4(i), and adjoining x_0 preserves compactness, proving (i).

(ii) If $\gamma_k \not\to 0$, there exist $\varepsilon > 0$ and infinitely many indices with $\gamma_k \ge \varepsilon$, so (82) follows immediately. Suppose therefore that $\gamma_k \to 0$. By (i) and Lemma 23, there exist $\overline{m}, \overline{L}^{\mathrm{d}}, C_{\mathcal{R}} > 0$ and k_0 such that, for all $k \ge k_0$,

$$m_k^\eta \le \overline{m}, \qquad \widehat{L}_k^{\mathrm{d},\eta} \le \overline{L}^{\mathrm{d}}, \qquad \left|\mathcal{R}^\eta_{k,\bar\beta_{k+1}} - 1\right| \le C_{\mathcal{R}}\gamma_k.$$

Equation (78) also gives

$$\left|\bar\beta_{k+1}\mathcal{R}^\eta_{k,\bar\beta_{k+1}} - 1\right| \le \bar\beta_{k+1}\left|\mathcal{R}^\eta_{k,\bar\beta_{k+1}} - 1\right| + \left|\bar\beta_{k+1} - 1\right| \le (\beta_{\mathrm{cap}}C_{\mathcal{R}} + s)\gamma_k. \tag{84}$$

Denote half the expression inside $[\cdot]_+$ in the denominator of the second candidate in (7) by

$$A_k := \frac{\bar\beta_{k+1}}{2}\mathcal{R}^\eta_{k,\bar\beta_{k+1}} - \frac{1}{2}\left(1 - \gamma_k\widehat{L}_k^{\mathrm{d},\eta} - \frac{\eta}{m_k^\eta}\right).$$

By the definition of A_k,

$$A_k - \frac{\eta}{2m_k^\eta} = \frac{1}{2}\left(\bar\beta_{k+1}\mathcal{R}^\eta_{k,\bar\beta_{k+1}} - 1\right) + \frac{1}{2}\gamma_k\widehat{L}_k^{\mathrm{d},\eta}.$$

Thus (84) yields

$$\left|A_k - \frac{\eta}{2m_k^\eta}\right| \le \frac{\eta}{2\overline{m}}C_0\gamma_k \le \frac{\eta}{2m_k^\eta}C_0\gamma_k, \qquad C_0 := \frac{\overline{m}}{\eta}\big(\beta_{\mathrm{cap}}C_{\mathcal{R}} + s + \overline{L}^{\mathrm{d}}\big).$$

Since $\gamma_k \to 0$, increase k_0 so that $C_0\gamma_k \le 1/2$. Then

$$0 < \frac{\eta}{2m_k^\eta}(1 - C_0\gamma_k) \le A_k \le \frac{\eta}{2m_k^\eta}(1 + C_0\gamma_k).$$

Using $1 - \beta_k/\bar\beta_k \ge 0$ from (64), the second candidate in (7) satisfies

$$\frac{\dfrac{\zeta_k^\eta}{1+\zeta_k^\eta}\left(1 - \dfrac{\beta_k}{\bar\beta_k}\right) + \dfrac{\eta}{2m_k^\eta}}{A_k} \ge \frac{1}{1 + C_0\gamma_k} \ge 1 - C_0\gamma_k. \tag{85}$$

The last inequality uses $1/(1+t) \ge 1 - t$ for $t \ge 0$. Since (78) gives $\bar\beta_{k+1} \ge 1$, (7) and (85) imply

$$\beta_{k+1} \ge 1 - C_0\gamma_k.$$

Since $C_0\gamma_k \le 1/2$, we obtain

$$\gamma_{k+1} \ge \gamma_k(1 - C_0\gamma_k) > 0.$$

Hence

$$\frac{1}{\gamma_{k+1}} \le \frac{1}{\gamma_k(1 - C_0\gamma_k)} \le \frac{1}{\gamma_k} + 2C_0, \tag{86}$$

where the last inequality uses $(1-t)^{-1} \le 1 + 2t$ for $0 \le t \le 1/2$. Summing (86) from k_0 to $k-1$ gives

$$\gamma_k \ge \frac{1}{\gamma_{k_0}^{-1} + 2C_0(k - k_0)}, \qquad k \ge k_0.$$

The series on the right diverges as a harmonic series, proving (82).

(iii) Use (74) from Lemma 22 and Proposition 24(ii). If $\liminf_{k\to\infty} G_k > 0$, there exist $\varepsilon > 0$ and k_1 such that $G_{k-1} \ge \varepsilon$ for all $k \ge k_1$. Therefore,

$$\sum_{k=k_1}^{\infty} \gamma_k G_{k-1} \ge \varepsilon \sum_{k=k_1}^{\infty} \gamma_k = \infty$$

contradicting (74). This proves (83). □

Finally, we prove stationarity along the entire sequence. For convex objectives, convergence of the Lyapunov function gave $x_k \to x_\star$. For weakly convex objectives, convergence to a global minimizer cannot generally be expected. We instead establish $\phi(\nabla f(x_k)) \to \phi(0)$ and hence $\nabla f(x_k) \to 0$.

C.5 Convergence to stationarity

Proof of Theorem 11. Let $\mathcal{C}$ be the compact set in Proposition 24(i). By Assumption 4 with $\eta > 0$, $\phi \circ \nabla f \in C^1(\mathbb{R}^d)$, so it is Lipschitz continuous on $\mathrm{conv}(\mathcal{C})$. Since $\nabla\phi(\nabla f(\cdot))$ is bounded on $\mathcal{C}$, there exist $L_P, M_{\mathcal{C}} > 0$ such that

$$|\phi(\nabla f(x)) - \phi(\nabla f(y))| \le L_P \|x - y\|, \qquad \|\nabla\phi(\nabla f(x))\| \le M_{\mathcal{C}}, \qquad x, y \in \mathcal{C}.$$

The update in Algorithm 1 gives

$$|G_{k+1} - G_k| \le L_P \|x_{k+1} - x_k\| = L_P \gamma_{k+1} \|\nabla\phi(\nabla f(x_k))\| \le L_P M_{\mathcal{C}} \gamma_{k+1}. \tag{87}$$

By Proposition 24(iii), $\liminf_{k\to\infty} G_k = 0$. Suppose that $G_k \not\to 0$. Then, for some $\varepsilon > 0$, we can choose infinitely many pairwise disjoint intervals $[a_j, b_j]$ such that $G_{a_j} \ge \varepsilon$ and $b_j > a_j$ is the first subsequent index with $G_{b_j} \le \varepsilon/2$. For $a_j \le k < b_j$, $G_k > \varepsilon/2$, and (87) gives

$$\frac{\varepsilon}{2} \le G_{a_j} - G_{b_j} \le \sum_{k=a_j}^{b_j-1} |G_{k+1} - G_k| \le L_P M_{\mathcal{C}} \sum_{k=a_j}^{b_j-1} \gamma_{k+1}.$$

Hence

$$\sum_{k=a_j+1}^{b_j} \gamma_k G_{k-1} \ge \frac{\varepsilon}{2} \sum_{k=a_j+1}^{b_j} \gamma_k \ge \frac{\varepsilon^2}{4 L_P M_{\mathcal{C}}}.$$

Each disjoint interval contributes a fixed positive amount to the left-hand side of (74), contradicting (74). Thus $G_k \to 0$.

Finally, Proposition 24(i) and continuity of ∇f imply that $\{\nabla f(x_k)\}$ lies in a compact set. For any convergent subsequence $\nabla f(x_{k_j}) \to y$, $G_k \to 0$ and continuity of ϕ give

$$\phi(y) = \phi(0).$$

By Assumption 4(ii), $y = 0$. Every cluster point of $\{\nabla f(x_k)\}$ is therefore 0, proving (17). □

D Preconditioner Design: Properties, Proofs, and Examples

We verify the properties of the construction in Section 5, derive the exponents for the principal target models, and collect the resulting preconditioners and their relation to existing methods in Table 3.

D.1 Proof of Proposition 12

Proof of Proposition 12. Fix a, b and suppress these parameters in the notation. Differentiating (19), with $\psi = \psi_{a,b}$ defined in (18), gives, for $\varrho > 0$,

$$\nu_{\mathrm{N}}''(\varrho) = \psi(\varrho), \qquad \frac{\nu_{\mathrm{N}}'(\varrho)}{\varrho} = \frac{1}{\varrho}\int_0^{\varrho} \psi(s)\, ds,$$

$$\frac{\nu_{\mathrm{T}}'(\varrho)}{\varrho} = \psi(\varrho), \qquad \nu_{\mathrm{T}}''(\varrho) = \psi(\varrho)\frac{1 + (a+1)\varrho^b}{1 + \varrho^b}.$$

For either construction allowed in the proposition, write $\nu = \nu_{\mathrm{N}}$ or $\nu = \nu_{\mathrm{T}}$ and $\phi(y) = \nu(\|y\|)$. Since $\psi > 0$, $\psi(0) = 1$, and $a \ge -1$ in the tangential case, both $\nu''(\varrho)$ and $\nu'(\varrho)/\varrho$ are positive and tend to 1 as $\varrho \to 0$. Also $\nu(0) = \nu'(0) = 0$, so $\nu(\varrho) = \varrho^2/2 + o(\varrho^2)$ and $\nabla\phi(0) = 0$. For $y \ne 0$ and $\varrho = \|y\|$,

$$\nabla\phi(y) = \frac{\nu'(\varrho)}{\varrho} y, \qquad \nabla^2\phi(y) = \nu''(\varrho)\frac{yy^\top}{\varrho^2} + \frac{\nu'(\varrho)}{\varrho}\left(I - \frac{yy^\top}{\varrho^2}\right).$$

Thus $\nabla\phi(y) = y + o(\|y\|)$, proving $\nabla^2\phi(0) = I$, and the Hessian formula shows that $\nabla^2\phi(y) \to I$ as $y \to 0$. Consequently, $\phi \in C^2(\mathbb{R}^d)$ and $\nabla^2\phi \succ 0$ everywhere. Since $\nu'(\varrho) > 0$ for $\varrho > 0$, its unique minimizer is 0. Moreover, ϕ is finite and convex on $\mathbb{R}^d$, so $\Omega_\phi = \mathbb{R}^d$ contains $\nabla f(\mathbb{R}^d)$ and

$\phi \in \mathcal{P}$. Its C^2 regularity gives local Lipschitz continuity of $\nabla\phi$, proving (i). Combined with the assumed properties of f, this also proves (iii).

For (ii), suppose that $\eta = 0$ and Assumption 4(i) holds, and fix a nonempty compact set $\mathcal{K} \subset \mathbb{R}^d$. Choose compact convex sets $\mathcal{C} \supset \mathcal{K}$ and $\mathcal{K}^* \supset \nabla f(\mathcal{C})$. Applying Lemma 16 to f on $\mathcal{C}$ and ϕ on $\mathcal{K}^*$ gives constants $m_f, M_f, M_\phi > 0$ such that, for all $x, y \in \mathcal{K}$,

$$D_\phi(\nabla f(x), \nabla f(y)) \le \frac{M_\phi}{2}\|\nabla f(x) - \nabla f(y)\|^2 \le \frac{M_\phi M_f^2}{2}\|x - y\|^2 \le \frac{M_\phi M_f^2}{m_f} D_f(y, x).$$

This proves Assumption 4(iii), and hence Assumption 4. The same lemma applied to ϕ on every nonempty compact convex subset of $\Omega_\phi = \mathbb{R}^d$ proves Assumption 8. □

D.2 Target models and the exponent a

For the power-law target $\tau(r) = r^p/p$ with $p > 1$, set $q = p/(p-1)$. Since $\varrho = \tau'(r) = r^{p-1}$, both inverse curvatures are proportional to ϱ^{q-2}, giving $a = q - 2 = (2-p)/(p-1)$. These targets cover $a > -1$, with growth exponent $p = (a+2)/(a+1)$.

For the exponential target $\tau(r) = e^r$, we have $\varrho = e^r$ and

$$h_{\mathrm{N}}(\varrho) = \frac{1}{\varrho}, \qquad h_{\mathrm{T}}(\varrho) = \frac{\log \varrho}{\varrho}. \tag{88}$$

Thus the normal inverse curvature is a power law with $a = -1$. For the tangential construction, we also choose $a = -1$, retaining the factor ϱ^{-1} and omitting $\log \varrho$. The exponent a reflects the target's inverse curvature, while $b > 0$ controls the interpolation with curvature 1 near the origin.

D.3 Representative preconditioners and existing methods

Table 3 gives the full set of examples, including the additional power-law variants omitted from Table 2. These illustrate the range of the construction; the empirical comparison of four preconditioners is given in Appendix E.4.

For the power-law target $\tau(r) = r^p/p$, the normal construction with $b = 2$ retains a one-dimensional integral for general q; the other listed finite-b examples have elementary expressions. The equality $\phi_{\mathrm{N}}^{-2,1} = \phi_{\mathrm{T}}^{-1,1}$ also shows that different curvature directions and exponents can yield the same preconditioner.

For the exponential target, the normal constructions additionally recover the tangential logarithmic factor in (88): their tangential curvatures are $\log(1+\varrho)/\varrho$ and $\operatorname{arsinh}(\varrho)/\varrho$, both asymptotic to $\log\varrho/\varrho$. In particular, directly interpolating the target's tangential inverse curvature by $\nu'(\varrho)/\varrho = \log(1+\varrho)/\varrho$ recovers $\phi_{\mathrm{N}}^{-1,1}$. This illustrates interpolation beyond pure power laws without introducing a new preconditioner. By contrast, the two tangential constructions with $a = -1$ have tangential curvature asymptotic to ϱ^{-1}, without the logarithmic factor.

In the clipping limit, $\psi_{-1,b}(\varrho) \to 1/\max\{1, \varrho\}$, so the update direction is $y/\max\{1, \|y\|\}$. The limiting function is linear in ϱ for $\varrho > 1$ and loses positive normal curvature; it is therefore outside Proposition 12.

The same function of the radius also defines a separable preconditioner $y \mapsto \sum_{i=1}^d \nu(|y_i|)$. Applying this to the NGD, HGD, and AdaGrad-type entries gives SNGD, SHGD, and memoryless coordinatewise AdaGrad-type updates (Oikonomidis et al., 2025a;b; Duchi et al., 2011). The AdaGrad-type specialization uses only the current squared gradient, with regularization 1 inside the square root; the radial entry uses its norm-based counterpart (Oikonomidis et al., 2025a, Example 1.5 and Lemma 1.3). The separable NGD construction also corresponds to Adam with both exponential decay parameters set to zero and denominator regularization 1 (Kingma & Ba, 2015).

E Experimental Details and Additional Results

We give the preprocessing, initial points, problem regularity, computational environment, and additional results.

Table 3: Preconditioning functions and related methods, where $q = p/(p-1) > 1$.

Preconditioning function	Relation
$\phi_{\mathrm{N}}^{0,b}(y) = \phi_{\mathrm{T}}^{0,b}(y) = \dfrac{\|y\|^2}{2} \quad (b > 0)$	GD
$\phi_{\mathrm{N}}^{q-2,1}(y) = \dfrac{(1+\|y\|)^q - 1 - q\|y\|}{q(q-1)}$	New variant
$\phi_{\mathrm{N}}^{q-2,2}(y) = \displaystyle\int_0^{\|y\|} (\|y\| - s)(1+s^2)^{(q-2)/2}\, ds$	New variant
$\phi_{\mathrm{T}}^{q-2,1}(y) = \dfrac{(1+\|y\|)^q - 1}{q} - \dfrac{(1+\|y\|)^{q-1} - 1}{q-1}$	New variant
$\phi_{\mathrm{T}}^{q-2,2}(y) = \dfrac{(1+\|y\|^2)^{q/2} - 1}{q}$	p-norm DPGD (Maddison et al., 2021)
$\phi_{\mathrm{N}}^{-1,1}(y) = (1+\|y\|)\log(1+\|y\|) - \|y\|$	Oikonomidis et al. (2025a, Table 1)
$\phi_{\mathrm{N}}^{-1,2}(y) = \|y\|\,\mathrm{arsinh}(\|y\|) - \sqrt{1+\|y\|^2} + 1$	HGD (Oikonomidis et al., 2025a;b)
$\phi_{\mathrm{T}}^{-1,1}(y) = \phi_{\mathrm{N}}^{-2,1}(y) = \|y\| - \log(1+\|y\|)$	NGD; exponential-penalty DPGD (Zhang et al., 2020; Maddison et al., 2021)
$\phi_{\mathrm{T}}^{-1,2}(y) = \sqrt{1+\|y\|^2} - 1$	AdaGrad-type (Duchi et al., 2011)
$\lim_{b\to\infty} \phi_{\mathrm{T}}^{-1,b}(y) = \begin{cases} \|y\|^2/2, & \|y\| \le 1, \\ \|y\| - 1/2, & \|y\| > 1 \end{cases}$	Gradient clipping (Zhang et al., 2020)

E.1 Data, preprocessing, and initial points

All experiments use the full distributed datasets; sample counts n and optimization dimensions d are specified for each dataset. For `kin8nm`, we standardize the features and response without adding an intercept. For `covtype`, let $A \in \mathbb{R}^{n\times 54}$ denote the feature matrix from `covtype.libsvm.binary.scale`. We map the labels to $\{-1, 1\}$ and whiten the 54 features to 53 dimensions without an intercept: retain eigenvalues of $A^\top A/n$ exceeding 10^{-10} times the largest, form W from the corresponding eigenvectors divided by the square roots of their eigenvalues, and set $B = AW$. Regularization and initialization use these optimization coordinates. The distributed features of `a9a` and `svmguide1` are used without further standardization or an added intercept. For `LEDGAR` (Tuggener et al., 2020) and `SCOTUS`, we use the distributed LIBSVM TF–IDF features without further transformation or an intercept. Each distributed training file combines the original LexGLUE (Chalkidis et al., 2022) training and validation sets; test data are excluded. We define a binary problem by labeling the most frequent class as $+1$ and all others as -1, breaking ties by the smallest numeric label. This gives 3661 positive samples (label 47) for `LEDGAR` and 1371 (label 0) for `SCOTUS`. These experiments compare optimization methods, not multiclass prediction accuracy. For `Skin Non-Skin` and `Rice Cammeo/Osmancik`, we standardize the features and append an unpenalized intercept; no features were removed. Standardization uses the mean and population standard deviation over all samples.

We initialize $x_0 = 0$ for `kin8nm`, `Skin Non-Skin`, and `Rice Cammeo/Osmancik`, and $x_0 = \nabla f(0)/\|\nabla f(0)\|$ in the whitened coordinates for `covtype`. For all powered hinge problems, choose x_0 as a positive multiple of $\nabla f(0)$ such that $\max_i [1 - y_i a_i^\top x_0]_+ = 10$.

E.2 Problem definitions and regularity

For $p > 2$, both $t \mapsto |t|^p$ and $t \mapsto [t]_+^p$ are C^2. Thus both convex objectives in Section 6 are C^2 and satisfy $\nabla^2 f \succeq I/n$. Together with Proposition 12, this gives the convex convergence and rate assumptions.

For the weakly convex objective in Section 6, the exponential loss is convex and each Cauchy penalty term satisfies

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\left[\omega \log\left(1 + \frac{t^2}{\sigma^2}\right)\right] = \frac{2\omega(\sigma^2 - t^2)}{(\sigma^2 + t^2)^2} \geq -\frac{\omega}{4\sigma^2}.$$

Thus f is η-weakly convex with $\eta = \omega/(4\sigma^2) = 0.025$. For both datasets, the prepared matrix A has full column rank and both classes are present. The exponential-loss Hessian is positive definite, and the Cauchy penalty plus $\eta\|x\|^2/2$ has a positive semidefinite Hessian, so $\nabla^2 g_\eta \succ 0$. On each sublevel set, the Cauchy penalty bounds all nonintercept coordinates and the exponential loss from both classes bounds the intercept. Hence f is level bounded; smoothness and Proposition 12 give the remaining weakly convex assumptions.

E.3 Reference values, timing, and computational environment

For each convex problem, we select the method requiring the fewest gradient evaluations to reach the common stopping criterion, among the eleven compared methods. We continue the selected method for 20,000 additional iterations and use the lowest objective value obtained as f_{ref}. For the runtime comparisons, we use per-iteration medians of 22 runs for convex problems and 10 for weakly convex problems, with balanced execution order; runtimes include initialization but exclude reference-value computation.

After the first step, Armijo trials start at twice the previous accepted stepsize and are halved until acceptance, with no fixed upper cap. The local stepsize estimate in (5) requires an auxiliary function evaluation but no gradient at that point; this cost is included in runtimes.

We use an Apple M1 CPU (8 cores), 16 GB memory, and macOS 26.4.1 (arm64), with double-precision arithmetic and one BLAS/OpenMP thread. The run records specify Python 3.13.14, NumPy 2.5.1, SciPy 1.18.0, and pandas 3.0.5; convex runs also use scikit-learn 1.9.0. For power-law targets, $\phi_{\mathrm{N}}^{a,2}$ need not have an elementary expression; we evaluate it and its gradient numerically using SciPy's `hyp2f1` in the ℓ_p and powered hinge loss experiments. Other tested preconditioners use elementary formulas.

Numerical safeguards use a 10^{-15} floor in Bregman ratios, clip negative Bregman values to zero, and select the growth candidate when curvature denominators are at most 10^{-15}. Removing these safeguards in 13 tested configurations left trajectories unchanged through the first relative-gradient crossing of 10^{-6}.

E.4 Additional numerical results

We use $\phi_{\mathrm{N}}^{a,1}, \phi_{\mathrm{N}}^{a,2}, \phi_{\mathrm{T}}^{a,1}, \phi_{\mathrm{T}}^{a,2}$ from Table 3, with a determined by each target. Each figure compares these four preconditioners under Adaptive DPGD and LS, giving eight curves. LS denotes DPGD linesearch (Maddison et al., 2021) for convex problems and preconditioned Armijo for weakly convex problems. Colors identify preconditioners, solid lines denote Adaptive DPGD, and dashed lines denote LS; distinct markers help distinguish overlapping curves. Holding the preconditioner fixed compares stepsize rules, while holding the rule fixed compares preconditioners. Convex plots show relative gradient norms and objective gaps against gradient evaluations and runtime; weakly convex plots show relative gradient norms against both horizontal axes. All curves retain the same gradient-based stopping criterion as the main comparison.

E.4.1 ℓ_p loss

For `kin8nm` from OpenML (Vanschoren et al., 2013) ($n = 8192$, $d = 8$), we set $p = 4$.

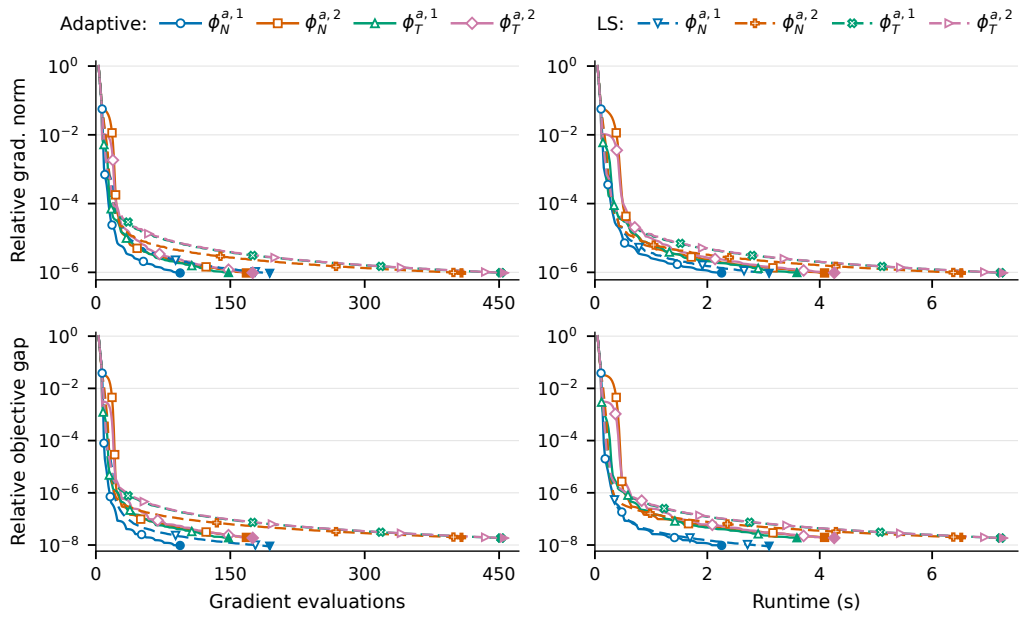


Figure 3: ℓ_6 loss on `covtype`. Rows: relative gradient norm and relative objective gap. Columns: gradient evaluations and runtime.

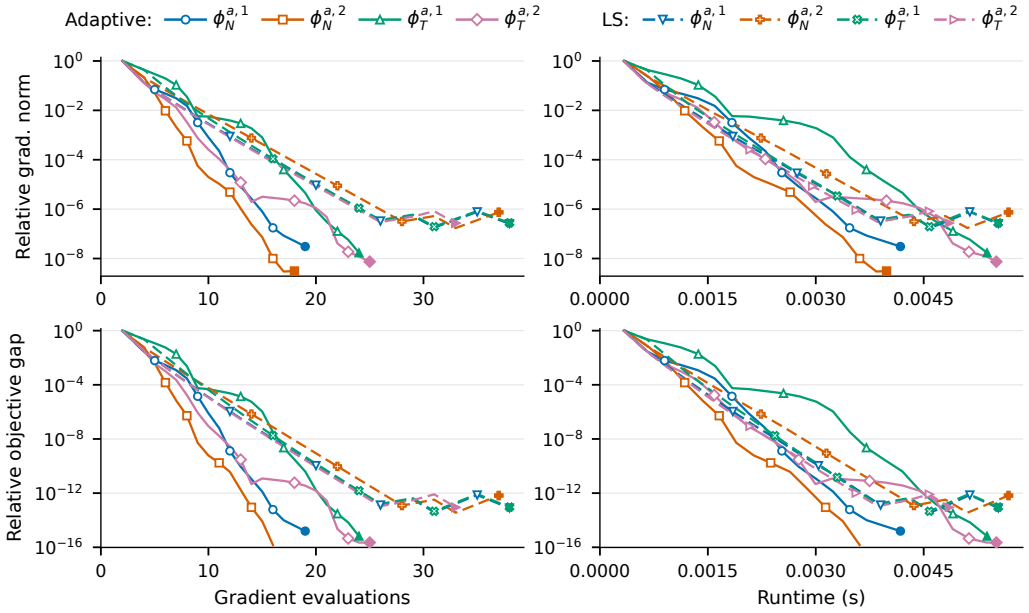


Figure 4: ℓ_4 loss on `kin8nm`. Rows and columns follow Figure 3.

E.4.2 Powered hinge loss

On `LEDGAR` and `SCOTUS`, Adaptive DPGD reaches the gradient stopping criterion sooner than LS with the same preconditioner (Figures 5 and 6). The four Adaptive variants are close in runtime. We also retain the full comparisons on `a9a` (Chang & Lin, 2011; Becker & Kohavi, 1996) ($n = 32561$, $d = 123$) and `svmguide1` (Chang & Lin, 2011) ($n = 3089$, $d = 4$). Adaptive DPGD reaches the gradient stopping criterion with fewer gradient evaluations and less runtime than LS for each preconditioner (Figures 7 and 8).

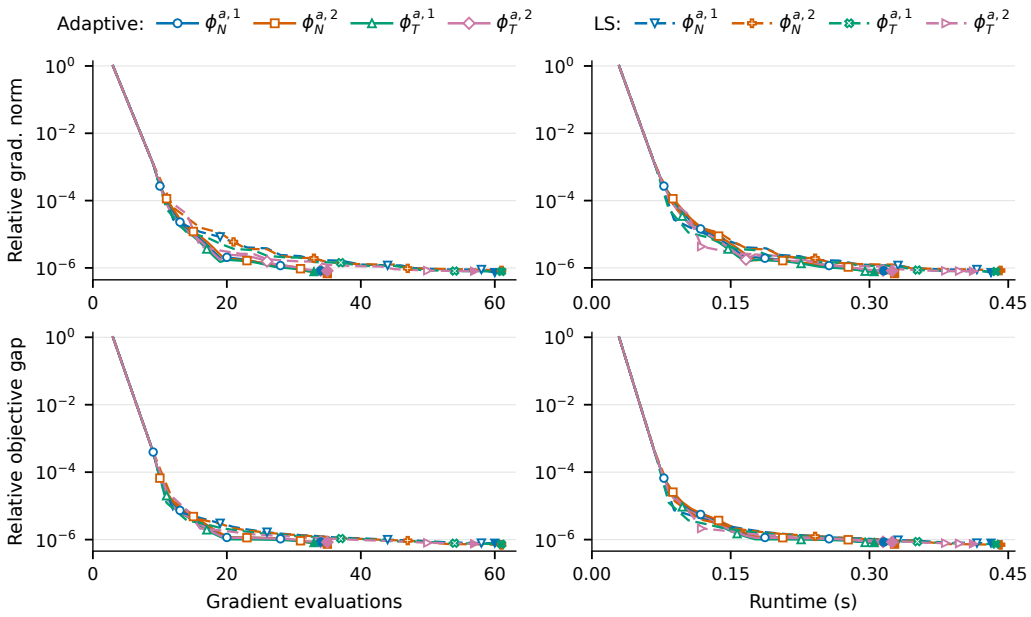


Figure 5: Powered hinge loss ($p = 6$) on LEDGAR. Rows and columns follow Figure 3.

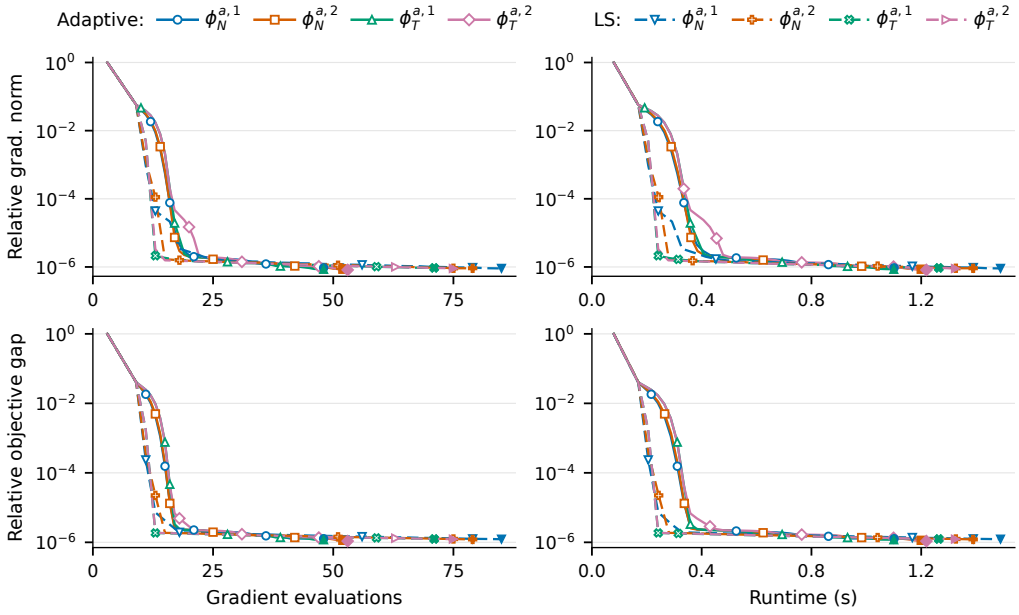


Figure 6: Powered hinge loss ($p = 6$) on SCOTUS. Rows and columns follow Figure 3.

E.4.3 Exponential loss with Cauchy penalty

On Skin Non-Skin, all Adaptive DPGD variants substantially outperform preconditioned Armijo linesearch (Figure 9). On Rice Cammeo/Osmancik, Adaptive DPGD variants other than $\phi_{\mathrm{T}}^{-1,1}$ substantially outperform preconditioned Armijo linesearch (Figure 10).

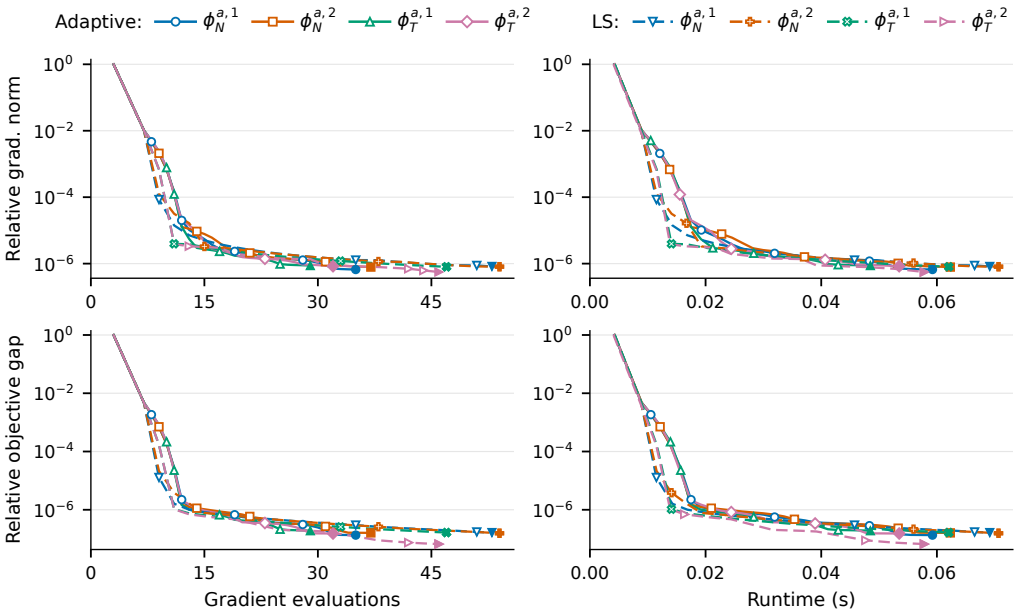


Figure 7: Powered hinge loss ($p = 6$) on `a9a`. Rows and columns follow Figure 3.

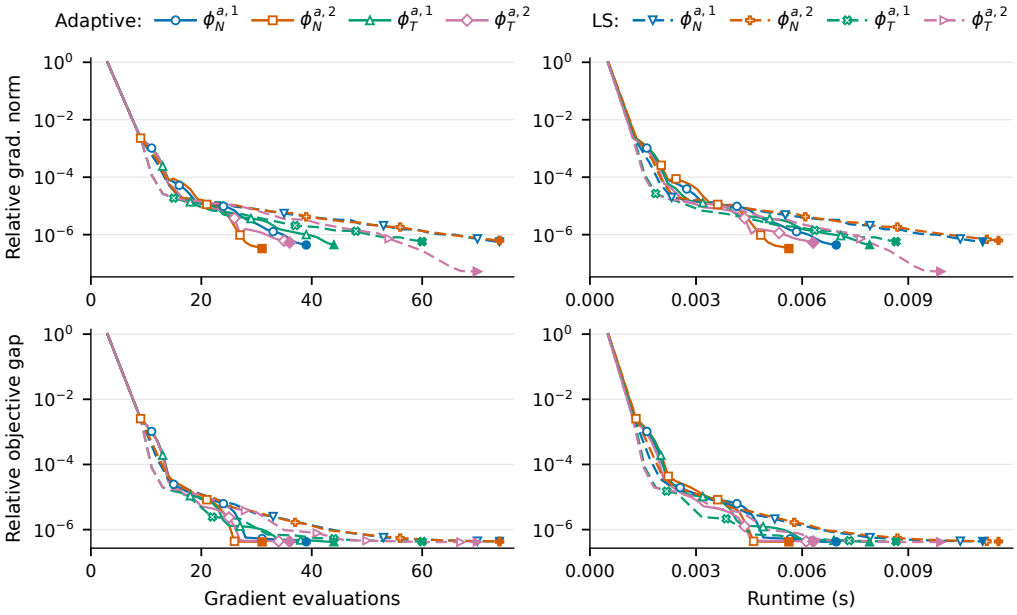


Figure 8: Powered hinge loss ($p = 6$) on `svmguide1`. Rows and columns follow Figure 3.

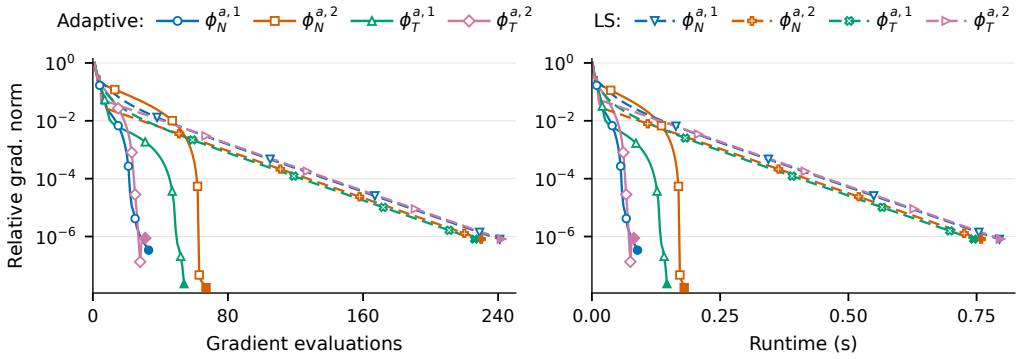


Figure 9: Exponential loss with Cauchy penalty on `Skin Non-Skin`. Relative gradient norms against gradient evaluations (left) and runtime (right).

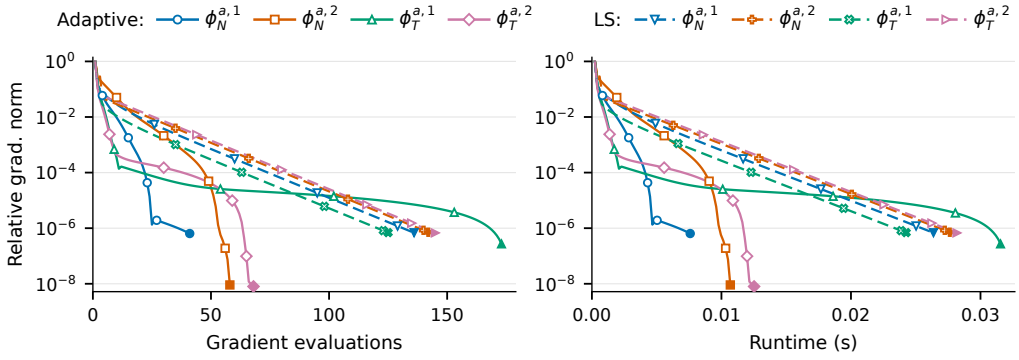


Figure 10: Exponential loss with Cauchy penalty on `Rice Cammeo/Osmancik`. Relative gradient norms against gradient evaluations (left) and runtime (right).

E.4.4 SUMMARY OF OBSERVATIONS

Under Adaptive DPGD, the p-norm regression preconditioner of Maddison et al. (2021) is $\phi_{\mathrm{T}}^{a,2}$, whereas $\phi_{\mathrm{N}}^{a,2}$ performs better on `kin8nm` (Figure 4) and $\phi_{\mathrm{N}}^{a,1}$ on `covtype` (Figure 3). Likewise, $\phi_{\mathrm{N}}^{-1,1}$ outperforms their exponential-penalty preconditioner $\phi_{\mathrm{T}}^{-1,1}$ on both weakly convex problems (Figures 9 and 10). Together with the powered hinge results, these observations illustrate the practical value of the design freedom allowed by our local conditions: alternative preconditioners can be constructed and applied across different losses without establishing a problem-specific global dual relative smoothness bound.

The numerical evaluation of $\phi_{\mathrm{N}}^{a,2}$ through `hyp2f1` does not dominate runtime in the convex experiments. With Adaptive DPGD, all evaluations of this preconditioner together account for less than 9% of total runtime in each of the six convex problems (median share over 22 timing repeats). Its gains in gradient evaluations are retained in runtime on `kin8nm` and `svmguide1`, where it reaches the stopping criterion fastest among the four tested preconditioners under Adaptive DPGD.

Overall, the normal constructions $\phi_{\mathrm{N}}^{a,1}$ and $\phi_{\mathrm{N}}^{a,2}$ provide effective choices among the tested preconditioners, with good performance in both gradient evaluations and runtime across several losses.

The comparison with DPGD linesearch also shows the benefit of avoiding repeated trial evaluations. After initialization, Adaptive DPGD uses one gradient evaluation per update, whereas the linesearch of Maddison et al. (2021) evaluates gradients at trial points. On `covtype` and all four powered hinge datasets, Adaptive DPGD reaches the stopping criterion with fewer gradient evaluations and less runtime for each of the four preconditioners. For example, with $\phi_{\mathrm{N}}^{a,1}$ on `covtype`, linesearch takes fewer updates but uses 194 gradient evaluations and 3.10 seconds, compared with 94 and 2.25 seconds for Adaptive DPGD. This illustrates how a linesearch-free update can improve runtime even when more updates are needed.

E.5 FULL COMPARISONS FOR THE MAIN EXPERIMENTS

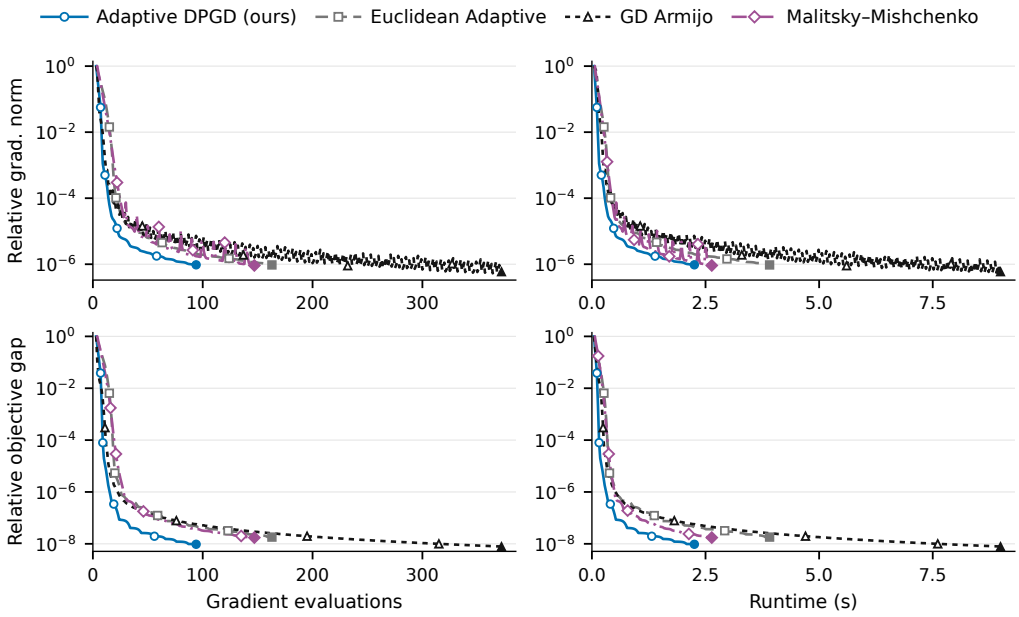


Figure 11: Full comparison for Figure 1(a) on `covtype`. Rows: relative gradient norm and relative objective gap. Columns: gradient evaluations and runtime.

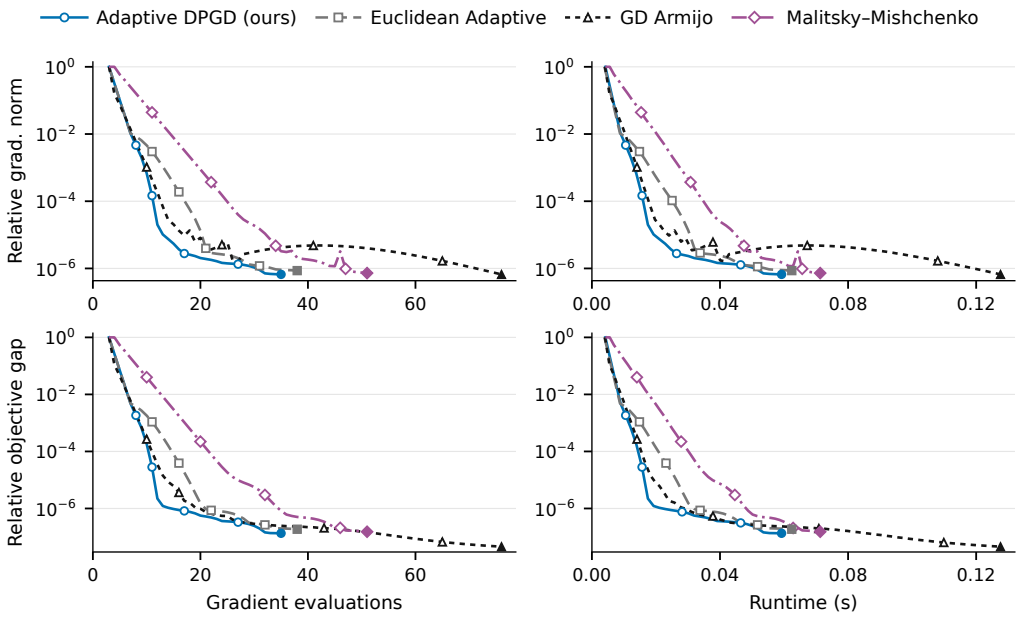


Figure 12: Full comparison for Figure 1(b) on `a9a`. Rows and columns follow Figure 11.

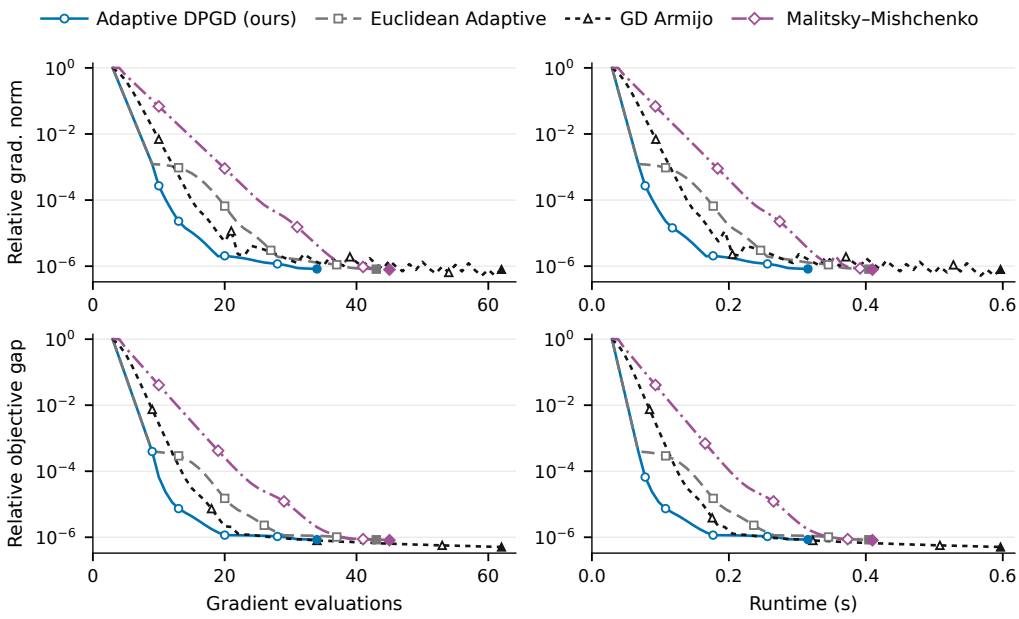


Figure 13: Full comparison for Figure 1(c) on LEDGAR. Rows and columns follow Figure 11.

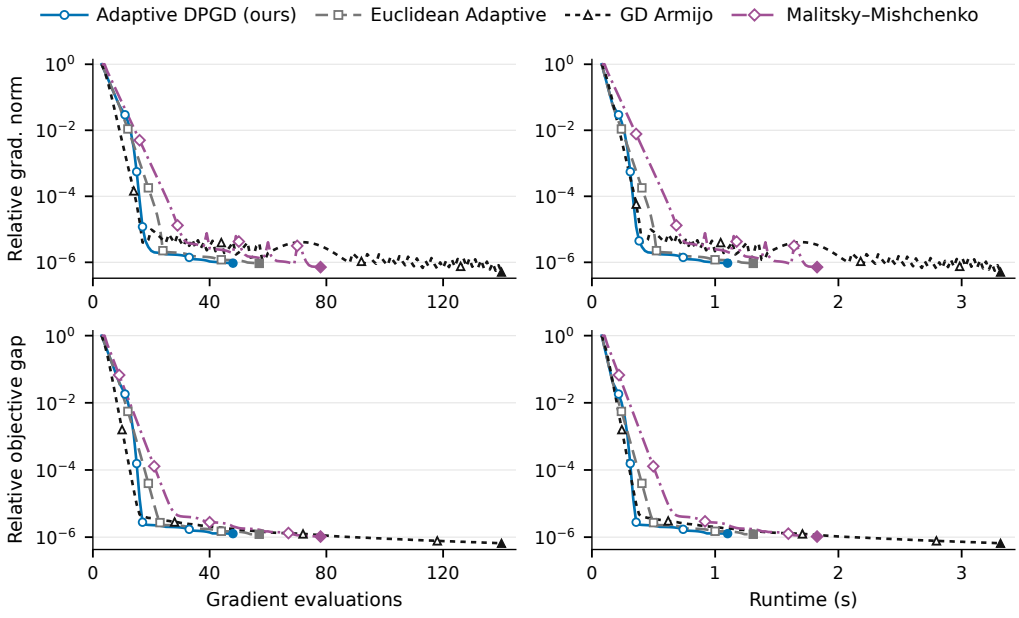


Figure 14: Full comparison for Figure 1(d) on SCOTUS. Rows and columns follow Figure 11.

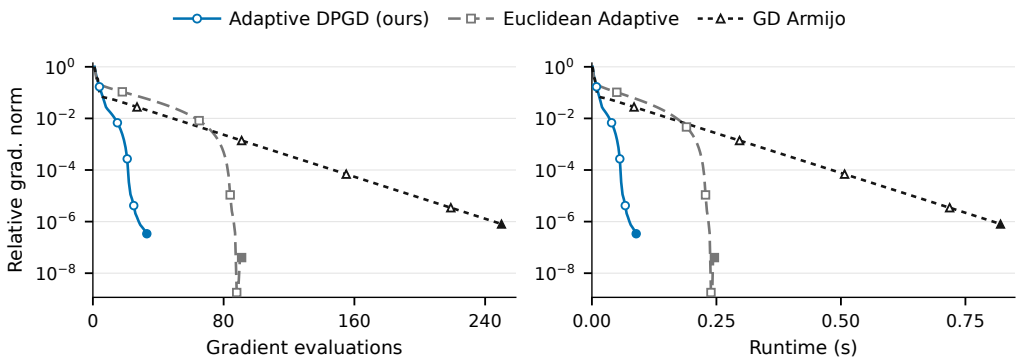


Figure 15: Full comparison for Figure 1(e) on `Skin Non-Skin`. Relative gradient norms against gradient evaluations (left) and runtime (right).

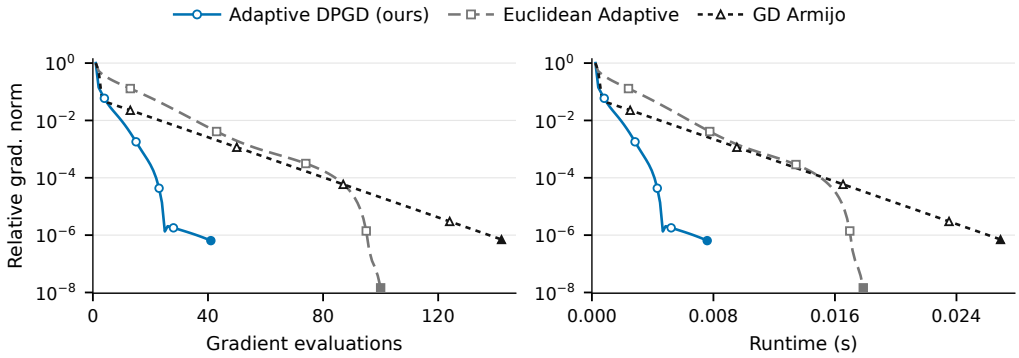


Figure 16: Full comparison for Figure 1(f) on `Rice Cammeo/Osmancik`. Relative gradient norms against gradient evaluations (left) and runtime (right).

F Extended Related Work

F.1 Non-Euclidean methods

Mirror descent (Nemirovski & Yudin, 1983) provides a classical framework for non-Euclidean geometry. Reference functions allow Bregman proximal gradient methods to control first-order model errors beyond global Lipschitz gradient continuity, with guarantees for convex (Bauschke et al., 2017), nonconvex (Bolte et al., 2018), and difference-of-convex problems (Takahashi et al., 2022). Quadratic approximations of Bregman divergences further simplify proximal subproblems (Takahashi & Takeda, 2025; Fujiki et al., 2025). For matrix factorization, reference functions tailored to the loss yield tractable simultaneous updates, including inertial Bregman proximal gradient methods (Mukkamala & Ochs, 2019) and majorization–minimization for nonnegative factorization with Kullback–Leibler loss (Takahashi et al., 2025). Bregman geometry also guides adaptive Frank–Wolfe stepsizes for weakly convex objectives (Takahashi et al., 2026).

For Legendre convex f, DPGD (Maddison et al., 2021) can be viewed as mirror descent on ϕ in gradient space, with f^* as the reference function. Here, the superscript $*$ denotes convex conjugation. It also relates preconditioning to inverse objective curvature: global dual relative smoothness compares $D_\phi(\nabla f(x), \nabla f(y))$ with $D_f(y, x)$. Its examples for p-norm regression and exponential penalties correct nonquadratic growth while remaining quadratic near the origin. Their guarantees require a global comparison for each objective–preconditioner pair. Our recipe makes this curvature-based construction systematic, while the local conditions in Proposition 12 permit the resulting preconditioners without a problem-specific global Hessian bound.

Oikonomidis et al. (2025a;b) unify normalization, clipping, and memoryless AdaGrad/Adam variants using a reference function φ, with $\nabla\varphi^*$ corresponding to our $\nabla\phi$. They motivate HGD through a nonlinear global upper bound based on $\varphi(y) = \cosh(\|y\|) - 1$. Our construction instead derives the same preconditioner from the normal inverse curvature of an exponential model, identifying the curvature it corrects and extending the procedure to other targets (Table 2).

The DPGD and nonlinear preconditioning guarantees above use global relative, anisotropic, or generalized smoothness conditions. DPGD linesearch also evaluates gradients at trial points. Adaptive DPGD uses local conditions and updates the stepsize directly from local curvature estimates, without a linesearch after initialization.

F.2 Further adaptive methods

Beyond the methods discussed in the main text, Yagishita & Ito (2025) study auto-conditioning based on local curvature estimates, and Hoai & Thai (2024) give a linesearch-free nonconvex proximal-gradient rule under additional structural assumptions. Patel & Berahas (2024) analyze nonconvex GD with diminishing stepsizes without global Lipschitz smoothness. Our focus is an explicit nonlinearly preconditioned update with adaptive stepsizes under local regularity and known weak convexity.

For convex optimization, Li & Lan (2026) develop a linesearch-free auto-conditioned fast gradient method with uniformly optimal rates across smoothness classes, while Suh & Ma (2025) propose a parameter-free adaptive Nesterov method without linesearch. Takahashi (2026) combines adaptive stepsizes with conditional gradient sliding for projection- and linesearch-free acceleration. Park et al. (2026) extend linesearch-free adaptive gradient descent to Riemannian manifolds, with applications to Gaussian variational inference.

Adaptive accelerated mirror descent (Xu & Chen, 2026) combines mirror descent and DPGD under global dual relative smoothness and additional curvature conditions, with backtracking when needed. Accelerating our local, linesearch-free methods remains a separate question.