IMPLICIT REGULARIZATION OF BREGMAN PROXIMAL POINT ALGORITHM ON SEPARABLE DATA: SHARP ALIGNMENT BOUNDS AND KULLBACK-LEIBLER GEOMETRIES

Anonymous authors
Paper under double-blind review

ABSTRACT

The Bregman proximal point algorithm (BPPA) has witnessed emerging machine learning applications such as knowledge distillation, policy optimization and fine-tuning, where the divergence is a design choice whose effect is not understood. We study this effect for linear classifiers trained on separable data with the exponential loss. The loss then vanishes along infinitely many directions, so the algorithm alone selects the classifier, which we judge by its margin. We first bound the margin by the alignment between the steps of the iterates and of their mirror images, which follow the negative gradient. If the alignment in a chosen norm stays above a constant and the divergence satisfies mild growth and curvature conditions, BPPA attains at least that constant times the maximum margin in that norm. The bound needs no strong convexity and recovers the condition-number bound of earlier work. It is tight: for every quadratic divergence and every norm, a one-point dataset attains the constant exactly. We then study two Kullback-Leibler (KL) geometries where the alignment has no positive lower bound, arising in exponentiated gradient and self-distillation. With KL on the parameters, for every constant stepsize the normalized ℓ_1-margin converges to its maximum, a discrete counterpart of a known mirror-flow result, whereas the Euclidean divergence selects the maximum ℓ_2-margin classifier. With KL on the predictions, when the samples are linearly independent and each proximal step is solved by gradient descent, BPPA converges in direction to the minimum-norm interpolator of the labels, a discrete analogue of a known natural-gradient-flow result. This limit maximizes the ℓ_2-margin if and only if every sample lies on the margin. Results for strongly convex and smooth divergences extend to mirror descent, and numerical experiments illustrate each result.

1 INTRODUCTION

The role of optimization methods has become arguably one of the most critical factors in the empirical performance of machine learning models (Kingma & Ba, 2014; Duchi et al., 2011; Goyal et al., 2017; Sutskever et al., 2013). Meanwhile, the Bregman proximal point algorithm (BPPA) (Eckstein, 1993; Kiwiel, 1997; Yang & Toh, 2021), a classical non-first-order method, has been drawing rising interest, with successes in knowledge distillation (Furlanello et al., 2018; Hinton et al., 2015), the mean-teacher paradigm (Tarvainen & Valpola, 2017), policy optimization (Schulman et al., 2015; Green et al., 2019) and fine-tuning of pre-trained models (Jiang et al., 2020). BPPA updates the parameters by minimizing a loss $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ while regularizing the distance to the previous iterate, measured by a divergence $\mathcal{D}$,

$$\theta_{t+1} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) + \frac{1}{2\eta_t} \mathcal{D}(\theta, \theta_t), \tag{1}$$

where $\{\eta_t\}$ are the stepsizes. The update suggests plausible intuitions for its success, such as damping aggressive updates and preventing catastrophic forgetting (Schulman et al., 2015; Li & Hoiem, 2018), but none of them has been rigorously justified, in sharp contrast with first-order methods.

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.

When the training objective has non-unique solutions, gradient descent is known to converge to solutions with favorable structure: the minimum-norm solution of under-determined linear systems (Gunasekar et al., 2018), the maximum-margin classifier on separable data (Soudry et al., 2018; Nacson et al., 2019), and margin-maximizing solutions of homogeneous networks (Lyu & Li, 2020). One would naturally ask:

Q1: Does BPPA converge to a structured solution with favorable properties?

Different divergences generate different iterates, and successful applications of BPPA hinge on a careful design of the divergence (Li & Hoiem, 2018; Hinton et al., 2015; Jiang et al., 2020), so understanding its role can reduce the effort of designing one for a new task. We therefore refine **Q1**:

Q2: How does the divergence affect the structure of the solution obtained by BPPA (and by mirror descent)?

We address **Q1** and **Q2** in the simplest setting in which the selection made by an optimizer can be characterized sharply (Soudry et al., 2018; Gunasekar et al., 2018): linear classifiers trained on linearly separable data with the exponential loss. The loss has infimum zero, approached only as $\|\theta\| \to \infty$ along an open cone of separators, so the algorithm alone selects the direction $\theta_t / \|\theta_t\|$, and we measure its quality by the *margin*, which controls generalization and robustness (Koltchinskii & Panchenko, 2002; Bartlett & Mendelson, 2002; Xu & Mannor, 2012). The applications above, with nonlinear models and often output divergences, motivate but do not instantiate this setting, except for output divergences on linear models (Section 5).

The main difficulty and our approach. The margin is measured in the primal space, but for a Bregman divergence $\mathcal{D} = D_w$ BPPA moves along the negative gradient in the dual (mirror) space of the distance-generating function (DGF) w: its optimality condition is $\nabla w(\theta_{t+1}) - \nabla w(\theta_t) = -2\eta_t \nabla \mathcal{L}(\theta_{t+1})$. For gradient descent the two displacements coincide, and the margin follows from a path-length argument (Soudry et al., 2018). For a general DGF they are not parallel, and this argument breaks. We measure the angle between them by the *alignment* c_t (Definition 3.1) and show a master inequality: up to lower-order terms, the normalized margin is at least the maximum margin times a path-weighted average of c_t (Proposition 3.2). For many divergences, the question of how the divergence shapes the solution then becomes how well its geometry keeps c_t away from zero. For the Kullback-Leibler (KL) divergence on the parameters and on the predictions, the alignment has no positive lower bound, and the solution is decided by additional structure of each geometry, which we analyze directly. Our contributions are as follows.

First, a positive alignment constant c_w yields a margin guarantee: under mild growth and curvature conditions, and without strong convexity, BPPA attains at least c_w times the maximum margin in any chosen norm (Theorem 3.5), which answers **Q1** and recovers the results of Li et al. (2021) with sharper finite-time bounds (Proposition 3.6).

Second, this constant is exact for quadratic DGFs: for every norm, some dataset attains the margin ratio c_w at every iteration (Theorem 3.8), so c_w is exactly the worst-case price of the divergence geometry, a quantitative answer to **Q2**.

Third, the divergence can also change the type of classifier. For the KL divergence on the parameters, $c_w = 0$, yet for every constant stepsize BPPA converges to the maximum ℓ_1-margin, whereas the Euclidean divergence selects the maximum ℓ_2-margin (Theorem 4.2). The proof for the discrete algorithm, whose continuous-time limit is known (Pesme et al., 2024), is our technical contribution.

Fourth, for the KL divergence on the predictions, as in self-distillation, BPPA with gradient-descent inner steps converges in direction to the minimum-norm interpolator of the labels when the samples are linearly independent (Theorem 5.3); this interpolator maximizes the ℓ_2-margin only when every sample lies on the margin. Numerical experiments compare each result with an exactly computable target (Section 6).

Relation to prior work. For separable data, gradient descent converges to the maximum ℓ_2-margin classifier (Soudry et al., 2018; Nacson et al., 2019; Ji & Telgarsky, 2021), and steepest descent to the maximum margin of its norm (Gunasekar et al., 2018). Mirror descent with homogeneous poten-

tials converges to the corresponding maximum margin (Sun et al., 2023), an analysis that excludes the entropy. Wang et al. (2023) obtain fast margin rates for strongly convex mirror geometries and view the assumption of Li et al. (2021) as essentially Euclidean, an assumption that Section 3 weakens. Multiplicative weights appear in these margin analyses on the dual (sample) side (Freund & Schapire, 1999; Ji & Telgarsky, 2021), whereas ours act on the primal coordinates. The hyperbolic entropy of Section 4 is also the mirror map of diagonal linear networks (Woodworth et al., 2020; Azulay et al., 2021).

2 Problem Setup

Notation. $[n] \coloneqq \{1, \dots, n\}$; for a norm $\|\cdot\|$ on $\mathbb{R}^d$, $\|\cdot\|_* = \max_{\|y\| \le 1} \langle \cdot, y \rangle$ is its dual norm; Δ_n is the probability simplex. For a dataset $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{+1, -1\}$, a linear classifier $u \in \mathbb{R}^d$ predicts $\operatorname{sign}(\langle u, x \rangle)$. We write $z_i \coloneqq y_i x_i$, let $Z \in \mathbb{R}^{n \times d}$ have rows $z_i^\top$, and assume throughout:

Condition 1 (Linear Separability)**.** *There exists $u \in \mathbb{R}^d$ such that $\langle u, z_i \rangle > 0$ for all $i \in [n]$.*

For a norm $\|\cdot\|$, the $\|\cdot\|_*$-norm margin of a separating classifier u is $\min_i \langle z_i, u/\|u\| \rangle$, the minimum distance from the data to the decision boundary measured in the $\|\cdot\|_*$-norm.

Definition 2.1 (Maximum $\|\cdot\|_*$-norm Margin)**.** *The maximum $\|\cdot\|_*$-norm margin classifier u^* and the maximum margin γ^* are*

$$u^* \in \operatorname*{argmax}_{\|u\| \le 1} \min_{i \in [n]} \langle u, z_i \rangle , \qquad \gamma^* \coloneqq \max_{\|u\| \le 1} \min_{i \in [n]} \langle u, z_i \rangle .$$

When $\|\cdot\|$ is strictly convex, u^* is unique. We write γ_2 and γ_1 for the maximum margins with $\|\cdot\| = \|\cdot\|_2$ and $\|\cdot\| = \|\cdot\|_1$, and $R_* \coloneqq \max_i \|x_i\|_*$ for the data radius. We minimize the empirical loss $\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(\langle \theta, z_i \rangle)$ with the exponential loss $\ell(x) = e^{-x}$. Under Condition 1, $\inf_\theta \mathcal{L}(\theta) = 0$ but no finite minimizer exists, so the object of study is the *limiting direction* $\theta_t / \|\theta_t\|$ and its margin.

For a convex and differentiable distance-generating function (DGF) $w : \mathbb{R}^d \to \mathbb{R}$, the Bregman divergence is $D_w(\theta, \theta') = w(\theta) - w(\theta') - \langle \nabla w(\theta'), \theta - \theta' \rangle$. BPPA (Eckstein, 1993; Kiwiel, 1997) and mirror descent (MD) (Nemirovsky & Yudin, 1983; Beck & Teboulle, 2003) instantiated by w update, from $\theta_0 = 0$,

$$\text{BPPA:} \quad \theta_{t+1} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) + \frac{1}{2\eta_t} D_w(\theta, \theta_t), \tag{2}$$

$$\text{MD:} \quad \theta_{t+1} = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \langle \nabla \mathcal{L}(\theta_t), \theta \rangle + \frac{1}{2\eta_t} D_w(\theta, \theta_t). \tag{3}$$

For $w = \|\cdot\|_2^2$, BPPA is the vanilla proximal point method and MD is gradient descent. The earlier analysis (Li et al., 2021) used the following condition.

Condition 2. *The DGF w is L_w-smooth and μ_w-strongly convex w.r.t. $\|\cdot\|$: $\frac{\mu_w}{2} \|\theta - \theta'\|^2 \le D_w(\theta, \theta') \le \frac{L_w}{2} \|\theta - \theta'\|^2$ for all $\theta, \theta' \in \mathbb{R}^d$.*

We write $\kappa_w \coloneqq L_w / \mu_w$ for the condition number of w relative to the *reference norm* $\|\cdot\|$, the norm in which the margin is evaluated, while the algorithm only sees w. Condition 2 covers every quadratic DGF $\theta^\top A \theta$ with $A \succ 0$, but it excludes the powers $\frac{1}{p} \|\theta\|_2^p$ and $\frac{1}{p} \|\theta\|_p^p$ for $p \neq 2$ and the entropy-type DGFs. Section 3 replaces it by a weaker condition, and Sections 4 and 5 treat geometries for which even the weaker condition fails.

3 The Alignment Constant and the Margin

This section develops the general tool: a master inequality for every DGF (Section 3.1), a margin guarantee under a positive alignment constant that needs no strong convexity (Section 3.2), the earlier results for strongly convex and smooth DGFs with sharper rates (Section 3.3), its exactness for quadratic DGFs (Section 3.4), and the geometries for which the constant vanishes (Section 3.5). Throughout, $\widetilde{\mathcal{O}}(\cdot)$ hides factors polylogarithmic in t and $1/\epsilon$.

3.1 The Alignment and a Master Inequality

The optimality condition of the proximal step (2) is

$$\nabla w(\theta_{t+1}) - \nabla w(\theta_t) = -2\eta_t \nabla \mathcal{L}(\theta_{t+1}), \tag{4}$$

so every BPPA step is a gradient step in the *dual* (mirror) space, taken at the new point. Write $\Delta_t := \theta_{t+1} - \theta_t$ for the primal displacement and $\Delta_t^* := \nabla w(\theta_{t+1}) - \nabla w(\theta_t)$ for the dual displacement. For gradient descent the two displacements are parallel, and for a general DGF we measure the angle between them.

Definition 3.1 (Alignment). *The alignment of the t-th step is $c_t := \langle \Delta_t^*, \Delta_t \rangle / (\|\Delta_t\| \, \|\Delta_t^*\|_*) \in [0, 1]$. With $r_t := \mathcal{L}(\theta_{t+1})/\mathcal{L}(\theta_t)$, its path-weighted average is $\bar{c}_T := \sum_{t<T} c_t r_t \|\Delta_t\| \,/\, \sum_{t<T} r_t \|\Delta_t\|$.*

The alignment measures how faithfully the geometry of w turns progress in the dual space, where the algorithm follows gradients, into progress in the primal space, where the margin is measured. For $w = \frac{1}{2}\|\cdot\|_2^2$ and the Euclidean norm, $c_t \equiv 1$.

Proposition 3.2 (Master inequality). *Let w be convex and differentiable, and assume the proximal step (2) is well posed. For any stepsizes and $T \geq 1$, with $P_T := \sum_{t<T} \|\Delta_t\|$ and $Q_T := \sum_{t<T}(1 - r_t)\|\Delta_t\|$,*

$$\min_{i\in[n]} \langle \theta_T, z_i \rangle \;\geq\; -\log n + \gamma^* \sum_{t<T} c_t \, r_t \, \|\Delta_t\| \;=\; -\log n + \gamma^* \, \bar{c}_T \, (P_T - Q_T). \tag{5}$$

Moreover $Q_T \leq R_* \sum_{t<T} \|\Delta_t\|^2$. If $\mathcal{L}(\theta_t) \to 0$ and $\|\Delta_t\| \to 0$, then

$$\liminf_{T\to\infty} \min_{i\in[n]} \langle \theta_T / \|\theta_T\|, z_i \rangle \;\geq\; \gamma^* \liminf_{T\to\infty} \bar{c}_T. \tag{6}$$

Proof idea. By convexity of $\mathcal{L}$, each step satisfies $\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t) \exp(-\langle \Delta_t^*, \Delta_t \rangle / (2\eta_t \mathcal{L}(\theta_t)))$. The optimality condition (4) and the bound $\|\nabla \mathcal{L}(\theta)\|_* \geq \gamma^* \mathcal{L}(\theta)$ give $\langle \Delta_t^*, \Delta_t \rangle = c_t \|\Delta_t\| \, \|\Delta_t^*\|_* \geq 2\eta_t \gamma^* \, c_t \, \|\Delta_t\| \, \mathcal{L}(\theta_{t+1})$, and summing over $t < T$ with $\mathcal{L}(\theta_T) \geq \frac{1}{n} e^{-\min_i \langle \theta_T, z_i \rangle}$ yields (5). Each unit of path length therefore raises the lower bound on the smallest logit by $\gamma^* c_t r_t$, and the geometry of w enters only through c_t (Appendix C). Every margin result of this section is (5) combined with a lower bound on c_t.

3.2 A Margin Guarantee from Uniform Alignment

The analysis of Li et al. (2021) assumes Condition 2, but its margin argument uses the condition only once: to show $c_t \geq \kappa_w^{-1/2}$ (Lemma B.3). By Proposition 3.2, any uniform lower bound on c_t would serve, which suggests the minimal requirement.

Condition 3 (Uniform alignment). *The DGF w is convex and differentiable, and*

$$c_w := \inf_{x \neq y} \frac{\langle \nabla w(x) - \nabla w(y), x - y \rangle}{\|x - y\| \, \|\nabla w(x) - \nabla w(y)\|_*} \;>\; 0.$$

Whether $c_w > 0$ does not depend on the reference norm, because all norms on $\mathbb{R}^d$ are equivalent, but its value does.

Proposition 3.3 (Sufficient conditions for uniform alignment). *(i) Under Condition 2, $c_w \geq \sqrt{\mu_w / L_w} = \kappa_w^{-1/2}$. (ii) If $w \in C^2$ and there are $\bar{\kappa} \geq 1$ and $\lambda : \mathbb{R}^d \to [0, \infty)$, positive outside a null set, with $\lambda(\theta)\|h\|^2 \leq \langle \nabla^2 w(\theta) h, h \rangle \leq \bar{\kappa}\, \lambda(\theta) \|h\|^2$ for all θ, h, then $c_w \geq \bar{\kappa}^{-1/2}$. If moreover $\|\cdot\| = \|\cdot\|_2$, then $c_w \geq 2\sqrt{\bar{\kappa}}/(1 + \bar{\kappa})$.*

Example 3.4 (Radial powers). *Case (ii) bounds only the condition number of the Hessian at each point, while its magnitude may vary freely. For $w(\theta) = \frac{1}{p}\|\theta\|_2^p$ with $p > 1$, the Hessian has eigenvalues $\|\theta\|_2^{p-2}$ and $(p-1)\|\theta\|_2^{p-2}$, so $c_w \geq 2\sqrt{p-1}/p$ in the Euclidean norm, while Condition 2 fails for every $p \neq 2$.*

Theorem 3.5 (Constant Stepsize BPPA under Uniform Alignment). *Assume Condition 3, $\theta_0 = 0$, the exponential loss, the growth condition $D_w(su^*, 0) \le C_G(1+s)^\rho$ for all $s \ge 0$, and one of the two curvature conditions (a) $\langle \nabla w(x) - \nabla w(y), x - y\rangle \ge \lambda_0(1+\max(\|x\|, \|y\|))^{-a} \|x-y\|^2$ for some $a < 2$, or (b) $D_w(x,y) \ge \frac{\mu_p}{p}\|x-y\|^p$ for some $p \ge 2$. Then for any constant stepsize $\eta > 0$:*

(1) $\mathcal{L}(\theta_t) \le \left[1 + C_G(1 + \log(2\eta t)/\gamma^*)^\rho\right]/(2\eta t)$ *for* $2\eta t \ge 1$;

(2) $\liminf_{t\to\infty} \min_i \langle \theta_t/\|\theta_t\|, z_i\rangle \ge c_w\gamma^*$;

(3) if $\sum_t \|\Delta_t\|^2 < \infty$, *which holds under (a) and under (b) with* $p < 3$, *then* $\min_i \langle \theta_T/\|\theta_T\|, z_i\rangle \ge c_w\gamma^* - \mathcal{O}(1/\log T)$.

Every DGF with a positive alignment constant that satisfies these growth and curvature conditions, strongly convex or not, therefore attains the margin $c_w\gamma^*$. The radial powers of Example 3.4 satisfy (a) with $a = 2 - p$ for $1 < p < 2$ and (b) for $p \ge 2$. Every DGF under Condition 2 satisfies all hypotheses with $\rho = 2$ and $a = 0$, and Section 3.3 states the guarantees under this condition.

Proof idea. In the earlier analysis, strong convexity bounds $\|\theta_t\|$ and the step lengths. Here the growth condition on $G(s) := D_w(su^*, 0)$ and the curvature condition take its place. Comparing θ_t with the scaled maximizer su^* through the three-point inequality gives $\mathcal{L}(\theta_t) \le e^{-s\gamma^*} + G(s)/(2\eta t)$ and $D_w(su^*, \theta_t) \le G(s) + 2\eta t\, e^{-s\gamma^*}$ for every $s \ge 0$. With $s = \log(2\eta t)/\gamma^*$ the first bound is (1). With $s = \log t/\gamma^*$ the second gives $D_w(su^*, \theta_t) \le G(s) + 2\eta = \mathcal{O}(\mathrm{polylog}\, t)$, and either curvature condition turns this Bregman bound into $\|\theta_t\| = \mathcal{O}(\mathrm{polylog}\, t)$ (Lemma C.11). The curvature condition and this norm bound then give $\|\Delta_t\| \to 0$ (Lemma C.12), so the defect Q_T in (5) is negligible and $c_t \ge c_w$ gives (2). Square-summable steps give the rate in (3).

3.3 Strongly Convex and Smooth DGFs: Earlier Results with Sharper Rates

Under Condition 2, Proposition 3.3(i) gives $c_w \ge \kappa_w^{-1/2}$, so Theorem 3.5 applies. The next proposition states the guarantees for BPPA and mirror descent under this condition, with T_ϵ the number of iterations after which the normalized margin is at least $(1-\epsilon)\kappa_w^{-1/2}\gamma^*$.

Proposition 3.6 (Strongly convex and smooth DGFs). *Assume Condition 2, $\theta_0 = 0$ and the exponential loss.*

(1) (constant stepsize) For every $\eta_t = \eta > 0$, $\mathcal{L}(\theta_t) = \mathcal{O}\big(\frac{L_w \log^2(e+\gamma^*\eta t)}{\gamma^{*2}\eta t}\big)$ *and* $\min_i \langle \theta_T/\|\theta_T\|, z_i\rangle \ge \kappa_w^{-1/2}\gamma^* - \mathcal{O}(1/\log(T+1))$, *so* $T_\epsilon \le \exp(\widetilde{\mathcal{O}}(1/\epsilon))$.

(2) (adaptive stepsize) For $\eta_t = \alpha_t/\mathcal{L}(\theta_t)$, *each step contracts the loss by a factor* $q(\alpha_t) < 1$ *that depends only on* $\alpha_t\gamma^{*2}/L_w$ *and tends to* 0 *as* $\alpha_t \to \infty$; *this part requires only the* L_w*-smoothness of* w. *With* $\alpha_t = 1/\sqrt{t+1}$, $\min_i \langle \theta_T/\|\theta_T\|, z_i\rangle \ge \kappa_w^{-1/2}\gamma^* - \mathcal{O}(\log(T+1)/\sqrt{T+1})$, *so* $T_\epsilon = \widetilde{\mathcal{O}}(1/\epsilon^2)$.

(3) (mirror descent) MD (3) satisfies (1) for $\eta \le \mu_w/(2R_*^2)$, *and the margin bound of (2) for* $\alpha_t = \min\{\mu_w/(2R_*^2), (t+1)^{-1/2}\}$.

The bound $\kappa_w^{-1/2}\gamma^*$ and the adaptive speed-up are due to Li et al. (2021), but the finite-time bounds are sharper: for BPPA the iteration count improves from $\exp(\widetilde{\mathcal{O}}(1/\epsilon^2))$ to $\exp(\widetilde{\mathcal{O}}(1/\epsilon))$ and from $\mathcal{O}(1/\epsilon^8)$ to $\widetilde{\mathcal{O}}(1/\epsilon^2)$ (MD and full statements: Table 1).

3.4 Quadratic DGFs: Exact Direction and Exact Constant

Quadratic DGFs satisfy Condition 2, and for them the guarantees above can be made exact. For $w(\theta) = \langle \theta, A\theta\rangle$ with $A \succ 0$ and the reference norm $\|\cdot\|_A := \sqrt{\langle \cdot, A\cdot\rangle}$, the DGF is 2-strongly convex and 2-smooth, so $\kappa_w = c_w = 1$.

Corollary 3.7 (Quadratic/Mahalanobis DGF). *With* $\|\cdot\| = \|\cdot\|_A$, $\theta_0 = 0$, *the exponential loss and any constant stepsize, BPPA with* $w(\theta) = \langle \theta, A\theta\rangle$ *converges to the maximum* $\|\cdot\|_{A^{-1}}$*-margin classifier:* $\lim_t \min_i \langle \theta_t/\|\theta_t\|, z_i\rangle = \gamma^*$ *and* $\lim_t \theta_t/\|\theta_t\| = u^*$.

Choosing A therefore steers BPPA toward a chosen Mahalanobis geometry. In another norm the margin can be smaller, and the next theorem shows that c_w is exactly the price.

Theorem 3.8 (Exact Sharpness). *Let $w(\theta) = \frac{1}{2}\theta^\top A\theta$ with $A \succ 0$ and let $\|\cdot\|$ be any norm. Then $c_w = \min_{\|\Delta\|=1} \langle A\Delta, \Delta\rangle / \|A\Delta\|_*$, attained at some Δ°. For the one-point dataset $\mathcal{S} = \{(A\Delta^\circ, +1)\}$, every constant stepsize $\eta > 0$ and $\theta_0 = 0$, the iterates satisfy $\theta_t = \rho_t\Delta^\circ$ with $\rho_t \uparrow \infty$, and*

$$\min_{(x,y)\in\mathcal{S}} \langle \theta_t / \|\theta_t\|, yx\rangle / \gamma^* = c_w \qquad \text{for every } t \geq 1.$$

In particular, $c_w = 2\sqrt{\kappa}/(1+\kappa)$ with $\kappa = \lambda_{\max}(A)/\lambda_{\min}(A)$ when $\|\cdot\| = \|\cdot\|_2$, and $c_w = 2/(1+\sqrt{d})$ when $A = I$ and $\|\cdot\| = \|\cdot\|_1$ on $\mathbb{R}^d$.

The earlier analysis (Li et al., 2021) showed sharpness of $\kappa_w^{-1/2}$ only up to a factor of two, along problems of growing dimension, whereas Theorem 3.8 gives equality in fixed dimension, and since $2\sqrt{\kappa}/(1+\kappa) > \kappa^{-1/2}$ for $\kappa > 1$, the earlier bound was never attained. Mirror descent behaves in the same way (Corollary D.3 and Proposition D.5).

3.5 When the Alignment Constant Vanishes

Condition 3 fails, for every reference norm, for the hyperbolic entropy underlying the KL divergence (Section 4) and the separable powers $\frac{1}{p}\sum_i |\theta_i|^p$ with $p > 2$, whose curvatures across coordinates have an unbounded ratio (Example C.10), and for the prediction-space KL divergence (Section 5), whose DGF is flat along the null space of Z (Lemma 5.1). For these DGFs the master inequality (5) still holds but yields no margin guarantee: the average $\bar{c}_T$ may degenerate along the trajectory, and for the prediction-space divergence the steps Δ_t do not even vanish. Sections 4 and 5 therefore use additional structure of each geometry: Section 4 replaces the alignment by a multiplicative-weights potential that compares the iterates with a fixed maximizer, and Section 5 solves the dynamics exactly in the logits.

4 Parameter-Space KL: Convergence to the Maximum ℓ_1-Margin

A central non-Euclidean geometry is the negative entropy $\sum_i \theta_i \log \theta_i$, whose Bregman divergence is the KL divergence and whose mirror descent is exponentiated gradient (Kivinen & Warmuth, 1997). For the corresponding mirror flow, the ℓ_1 limit follows from the horizon-function analysis of Pesme et al. (2024), and this section treats the discrete algorithm with a finite stepsize.

Setup. Negative entropy lives on the positive orthant, while a separating classifier must range over $\mathbb{R}^d$ with $\|\theta_t\| \to \infty$. We therefore use the standard EG$^\pm$ lift (Kivinen & Warmuth, 1997; Ghai et al., 2020): write $\theta = a - b$ with $a, b \in \mathbb{R}^d_{++}$ and run BPPA on (a, b) with the unnormalized KL divergence $\mathrm{KL}(a\|a') = \sum_i \left(a_i \log \frac{a_i}{a'_i} - a_i + a'_i\right)$,

$$(a_{t+1}, b_{t+1}) = \operatorname*{argmin}_{a,b\in\mathbb{R}^d_{++}} \mathcal{L}(a-b) + \frac{1}{2\eta}\big[\mathrm{KL}(a\|a_t) + \mathrm{KL}(b\|b_t)\big], \tag{7}$$

from $a_0 = b_0 = \beta \in \mathbb{R}^d_{++}$, so that $\theta_0 = 0$. The natural reference norm is $\|\cdot\|_1$. We write $R_\infty := \max_i \|x_i\|_\infty$, u^* for a maximum ℓ_1-margin classifier ($\|u^*\|_1 = 1$), and $\mu_1(\theta) := \min_i \langle \theta/\|\theta\|_1, z_i\rangle$. By linear programming duality, $\gamma_1 = \min_{q\in\Delta_n} \left\|Z^\top q\right\|_\infty$.

Lemma 4.1 (KL-BPPA is hyperbolic-entropy BPPA). *The step (7) has a unique minimizer, and $a_{t,i}b_{t,i} = \beta_i^2$ for all t, i. With $\psi_t := \frac{1}{2}\log(a_t/b_t)$, one has $a_t = \beta e^{\psi_t}$, $b_t = \beta e^{-\psi_t}$, $\theta_t = 2\beta\sinh\psi_t$ componentwise, and $\psi_t = \nabla w_\beta(\theta_t)$ for the hyperbolic-entropy DGF*

$$w_\beta(\theta) := \sum_{i=1}^d \left[\theta_i \operatorname{arcsinh}\frac{\theta_i}{2\beta_i} - \sqrt{\theta_i^2 + 4\beta_i^2}\right], \qquad \nabla^2 w_\beta(\theta) = \operatorname{diag}\Big(\frac{1}{\sqrt{\theta_i^2+4\beta_i^2}}\Big).$$

Hence (θ_t) is BPPA (2) on $\mathbb{R}^d$ instantiated by w_β, and in dual form $\psi_{t+1} = \psi_t - 2\eta\nabla\mathcal{L}(\theta_{t+1})$.

The DGF w_β is smooth but not strongly convex, and its diagonal Hessian decays at coordinate-dependent rates, so $c_{w_\beta} = 0$ for every norm (Example C.10) and Section 3 yields no margin guarantee. The next theorem is proved by a different argument.

Theorem 4.2 (KL-BPPA converges to the maximum ℓ_1-margin). *Let $\eta > 0$ be constant and $a_0 = b_0 = \beta$. Write $\tau_t \coloneqq 2\eta\mathcal{L}(\theta_{t+1})$ and $A_T \coloneqq \sum_{t<T} \tau_t$. Then:*

(0) (loss and norm) $\mathcal{L}(\theta_t) = \mathcal{O}(\log t \, \log\log t/t)$, and $\|\theta_t\|_1$ grows like $\log t$ up to $\log\log t$ factors;

(a) (integrated margin gap) $\sum_{t\geq 0} \tau_t\big(\gamma_1 - \mu_1(\theta_{t+1})\big) \leq C_\infty < \infty$, with C_∞ explicit in $n, \beta, \eta, \gamma_1, R_\infty$ (Appendix E);

(b) (support persistence) $\|\psi_T\|_\infty - \langle \psi_T, u^*\rangle \leq C_\infty$ for every T; hence $\operatorname{sign}(u_i^*)\,\psi_{T,i} \geq \|\psi_T\|_\infty - C_\infty/|u_i^*|$ for $i \in \operatorname{supp}(u^*)$, so, for all large T, every support coordinate of θ_T has the sign of u^* and is at least a constant fraction of $\|\theta_T\|_\infty$;

(c) (margin) $\mu_1(\theta_t) \to \gamma_1$;

(d) (direction) every limit point of $\theta_t/\|\theta_t\|_1$ is a maximum ℓ_1-margin classifier; if u^* is unique, $\theta_t/\|\theta_t\|_1 \to u^*$.

On the same data the Euclidean divergence leads to the maximum ℓ_2-margin classifier (Corollary 3.7) and the KL divergence to maximum ℓ_1-margin classifiers (Figure 1(b)). Parts (a)–(d) describe this bias as an integrated bound, on the support, on the margin and on the direction.

Proof idea. Let $q_t \in \Delta_n$ be the loss weights, $q_{t,i} \propto e^{-\langle \theta_t, z_i\rangle}$, so that $-\nabla\mathcal{L}(\theta) = \mathcal{L}(\theta) Z^\top q(\theta)$ and the dual update is $\psi_{t+1} = \psi_t + \tau_t Z^\top q_{t+1}$. The *dual clock* is $A_T = \sum_{t<T}\tau_t$. Since $a_{t+1,i} = a_{t,i} e^{\psi_{t+1,i} - \psi_{t,i}}$ and $b_{t+1,i} = b_{t,i} e^{-(\psi_{t+1,i} - \psi_{t,i})}$, the lifted weights form a multiplicative-weights (Hedge) learner over $2d$ experts with cumulative rewards $\pm\psi_{t,i}$, with potential $P_t \coloneqq \sum_i (a_{t,i} + b_{t,i}) \in [\|\theta_t\|_1, \|\theta_t\|_1 + 2\sum_i \beta_i]$. The proof has three steps. (i) *Two bounds on the potential.* Hoeffding's lemma bounds its growth, and the best expert bounds it from below because $\langle u^*, Z^\top q\rangle \geq \gamma_1$ for every $q \in \Delta_n$:

$$\gamma_1 A_T + \log\beta_{\min} \;\leq\; \log P_T \;\leq\; \log P_0 + \sum_{t<T} \frac{\tau_t \langle \theta_t, Z^\top q_{t+1}\rangle}{P_t} + \frac{{R_\infty}^2}{2}\sum_{t<T}\tau_t^2.$$

(ii) *Summable errors.* By the Gibbs variational identity, the reward per unit of dual clock is at most the current normalized margin plus an offset, $\langle \theta_t, Z^\top q_{t+1}\rangle / P_t \leq \mu_1(\theta_{t+1}) + \big(\log n + 2R_\infty(\|\Delta_t\|_1 + \sum_i \beta_i)\big)/P_t$. Combining with (i),

$$\sum_{t<T} \tau_t\big(\gamma_1 - \mu_1(\theta_{t+1})\big) \;\leq\; \log\frac{P_0}{\beta_{\min}} + C\sum_{t<T}\frac{\tau_t}{P_t} + 2R_\infty \sum_{t<T}\frac{\tau_t\|\Delta_t\|_1}{P_t} + \frac{{R_\infty}^2}{2}\sum_{t<T}\tau_t^2,$$

with $C = \log n + 2R_\infty\sum_i \beta_i$. The last two sums are finite by the loss rate in (0), since $\|\Delta_t\|_1 \leq c\,\tau_t P_t$. The first is finite because the lower bound in (i) gives $P_t \geq \beta_{\min} e^{\gamma_1 A_t}$, so $\sum_t \tau_t/P_t \leq \beta_{\min}^{-1}\sum_t \tau_t e^{-\gamma_1 A_t} < \infty$. This gives (a), and the two bounds in (i) give (b). (iii) *From integrated to pointwise.* The dual clock diverges and, for large t, $|\mu_1(\theta_{t+2}) - \mu_1(\theta_{t+1})| \leq C_L\tau_{t+1}$ with a constant C_L, so a gap $\gamma_1 - \mu_1 \geq \epsilon$ on infinitely many disjoint clock intervals of length $\epsilon/(2C_L)$ would contradict (a). This gives (c), and (d) follows by compactness. The full proof is in Appendix E.

Two remarks are in order. First, the argument explains why the discrete KL geometry escaped previous analyses: for gradient descent the primal and dual steps are parallel and the margin follows from the dual update alone (Soudry et al., 2018; Ji & Telgarsky, 2021), whereas here the primal step is the dual step reweighted coordinatewise by $\sqrt{\theta_{t,i}^2 + 4\beta_i^2}$ and the alignment is never controlled. Comparing with the *fixed* expert u^* avoids any convergence requirement on q_t. Second, Theorem 4.2 gives no rate: (a) controls the gap only on average over the dual clock, which grows like $\log\|\theta_T\|_1 \asymp \log\log t$, and a pointwise rate would require a rate for q_t.

5 Prediction-Space KL: Convergence to the Minimum-Norm Label Interpolator

The divergences used in distillation and self-training are defined on the *predictions* of the model rather than on its parameters (Hinton et al., 2015; Furlanello et al., 2018; Yuan et al., 2020). For

the linear logistic model $p_\theta(x) = \sigma(\langle \theta, x \rangle)$ with $\sigma(s) = (1 + e^{-s})^{-1}$, write $s_i(\theta) \coloneqq \langle \theta, z_i \rangle$ for the signed logits and $A(s) \coloneqq \log(1 + e^s)$ for the log-partition function, so that $A' = \sigma$. Since $\mathrm{KL}(\mathrm{Bern}(\sigma(s'))\|\mathrm{Bern}(\sigma(s))) = A(s) - A(s') - \sigma(s')(s - s')$ is the Bregman divergence of A, the prediction-space KL divergence $D_{\mathrm{pred}}(\theta, \theta') \coloneqq \frac{1}{n}\sum_i \mathrm{KL}(\mathrm{Bern}(p_{\theta'}(x_i)) \,\|\, \mathrm{Bern}(p_\theta(x_i)))$ satisfies

$$D_{\mathrm{pred}}(\theta, \theta') = D_W(\theta, \theta'), \quad W(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^{n} A\big(\langle \theta, z_i \rangle\big), \quad \nabla^2 W(\theta) = \frac{1}{n}\sum_{i=1}^{n} \sigma'(s_i(\theta))\, z_i z_i^\top. \tag{8}$$

For a linear model, D_{pred} is therefore a genuine but degenerate Bregman divergence over the parameter space, and since $z_i z_i^\top = x_i x_i^\top$, $\nabla^2 W$ is the Fisher information: the proximal step lives in the geometry of natural gradient. For its flow with the logistic loss and $n < d$, Kerekes et al. (2021) showed that, with the pseudo-inverse choice of direction, the parameters converge in direction to the least-squares interpolant of the labels, and Theorem 5.3 below is the discrete counterpart for the exponential loss, with a rate and an exact criterion for when the limit is the maximum ℓ_2-margin classifier. Since $c_W = 0$ (Lemma 5.1), the proof does not bound the alignment but solves the dynamics exactly in the logits, which decouple.

Lemma 5.1 (Structure of the prediction-space geometry). *(i) W is convex, and $D_W(\theta+v, \theta'+v') = D_W(\theta, \theta')$ for all $v, v' \in \mathrm{null}(Z)$. Consequently, if $\mathrm{rank} Z < d$, the proximal step $\mathrm{argmin}_\theta\, \mathcal{L}(\theta) + \frac{1}{2\eta} D_W(\theta, \theta_t)$ has an affine set of minimizers. (ii) $c_W = 0$ for every reference norm if $\mathrm{rank} Z < d$, or if $d \geq 2$ and no two of the z_i are parallel.*

Part (i) holds because W, ∇W and the loss depend on θ only through the logits $Z\theta$. For (ii), the weights $\sigma'(s_i) \asymp e^{-|s_i|}$ in (8) decay at sample-dependent rates along a direction $R\widehat{\theta}$ with $R \to \infty$, so the condition number of $\nabla^2 W$ diverges. By (i), when $\mathrm{rank} Z < d$ the proximal step must be complemented by a selection rule.

Lemma 5.2 (Logit reduction and the natural selection). *Let $\mathrm{rank} Z = n$, so that $n \leq d$, and let $Z^\dagger \coloneqq Z^\top K^{-1}$ with $K \coloneqq ZZ^\top$. The minimizers of the proximal step are exactly $\{\theta : Z\theta = s_{t+1}\}$, where each logit follows the same scalar map $s_{t+1,i} = T_\eta(s_{t,i})$; for the exponential loss, $T_\eta(s)$ is the unique solution s' of $\sigma(s') - \sigma(s) = 2\eta e^{-s'}$. Gradient descent on the proximal objective started at θ_t converges to the minimizer $\theta_t + Z^\dagger(s_{t+1} - s_t)$.*

We adopt this *natural selection* $\theta_{t+1} = \theta_t + Z^\dagger(s_{t+1} - s_t)$, the solution reached by Euclidean gradient descent on each proximal step started at θ_t, which plays the role of the pseudo-inverse choice of Kerekes et al. (2021). It keeps the null-space component of θ_0 fixed, since $\theta_t = Z^\dagger s_t + (I - Z^\dagger Z)\theta_0$.

Theorem 5.3 (Prediction-space KL BPPA, linearly independent samples). *Let $\mathrm{rank} Z = n$, the exponential loss, any constant $\eta > 0$, any θ_0, and the natural selection. Then:*

(1) (geometric decay) $\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t)/(1 + 2\eta)$ for all t;

(2) (equal logit growth) $s_{t,i} = s_{0,i} + t\log(1+2\eta) + v_i - \varepsilon_{t,i}$, with constants $v_i > 0$ and $0 < \varepsilon_{t,i} \leq v_i$, $\varepsilon_{t,i} = \mathcal{O}((1 + 2\eta)^{-t})$;

(3) (direction) $\left\| \theta_t / \|\theta_t\|_2 - Z^\dagger \mathbf{1} / \left\| Z^\dagger \mathbf{1} \right\|_2 \right\|_2 = \mathcal{O}(1/t)$, where $Z^\dagger \mathbf{1} = X^\dagger y$ is the minimum-norm solution of $X\theta = y$;

(4) (margin) $\lim_{t\to\infty} \min_i \langle \theta_t / \|\theta_t\|_2, z_i \rangle = (\mathbf{1}^\top K^{-1} \mathbf{1})^{-1/2} \leq \gamma_2$, with equality if and only if $K^{-1}\mathbf{1} \geq 0$ componentwise, that is, if and only if every sample is a support vector of the hard-margin SVM.

A few remarks are in order. First, the limit is the minimum-norm interpolator of the labels rather than a maximum-margin classifier. The mechanism is visible in (2): the proximal term forces every logit, including those of easy points, to grow at the same rate $\log(1 + 2\eta)$. For $z_1 = (1, 0)$ and $z_2 = (10, 0.1)$, $\gamma_2 = 1$, $K^{-1}\mathbf{1} = (9001, -900)$ and the limiting margin is $1/\sqrt{8101} \approx \gamma_2/90$, and from the SVM direction, the iteration walks away (Figure 1(d)). Second, here a support vector is a sample on the margin of the SVM. In high dimension every sample is one with high probability (Hsu et al., 2021; Muthukumar et al., 2021), and then the prediction-space proximal term is harmless for the margin. Third, the squared prediction distance $\frac{1}{2n}\|Z(\theta - \theta')\|_2^2$, LS for short, selects the same direction $Z^\dagger \mathbf{1}$ when $\mathrm{rank} Z = n$, but its loss decays only like $1/t$ (Proposition F.2).

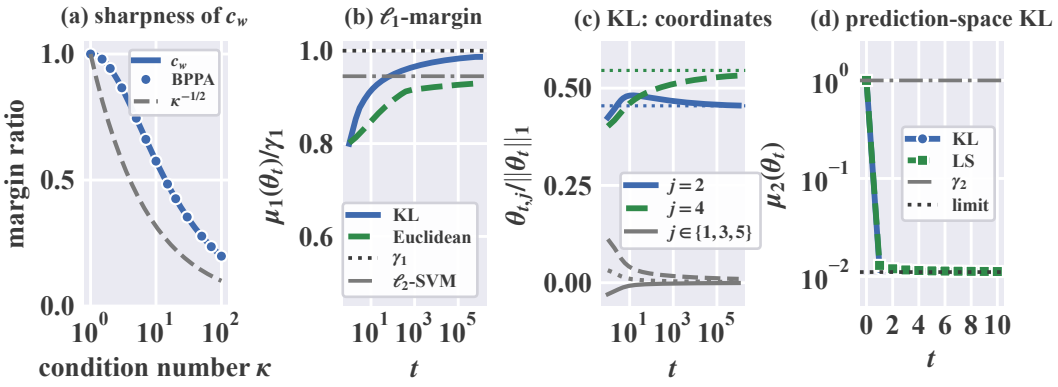

Figure 1: (a) Theorem 3.8: margin ratio on the worst-case one-point dataset. (b, c) Theorem 4.2: ℓ_1-margin, and coordinates of $\theta_t/\|\theta_t\|_1$ for KL-BPPA (dotted: u_2^*, u_4^*). (d) Theorem 5.3: ℓ_2-margin of KL and LS proximal steps, dotted: the limit $(\mathbf{1}^\top K^{-1}\mathbf{1})^{-1/2}$.

6 Numerical Illustrations

The experiments check the three main results on small datasets whose targets are exactly computable (Figure 1; datasets and further panels in Appendix G). Every proximal step is solved to machine precision, and every reference value is computed exactly. Appendices G and H add constant versus adaptive stepsizes (Proposition 3.6) and CIFAR-100 experiments with the KL and squared prediction distances.

Sharpness of the alignment constant. On the worst-case one-point dataset of Theorem 3.8 with $A = \mathrm{diag}(1, \kappa)$, the ℓ_2-margin ratio of BPPA equals $2\sqrt{\kappa}/(1+\kappa)$ at every iteration, up to 4×10^{-16}, strictly above the earlier guarantee $\kappa^{-1/2}$ (panel a), and for other one-point datasets it is larger (Figure 3).

Parameter-space KL. On three samples in $\mathbb{R}^5$ whose maximum ℓ_1-margin classifier is unique, $u^* = (0, \frac{5}{11}, 0, \frac{6}{11}, 0)$ with $\gamma_1 = 69/110$ by linear programming duality, KL-BPPA reaches $0.987\,\gamma_1$ after 10^6 steps, whereas Euclidean BPPA approaches the ℓ_1-margin of the ℓ_2-SVM, $0.945\,\gamma_1$ (panel b). The support coordinates stay bounded away from zero and the off-support coordinates decay, as Theorem 4.2(b) and (d) predict (panel c). The convergence is slow, consistently with the absence of a rate (Figure 5).

Prediction-space KL. On the two-point example of Section 5, started from the SVM direction with $\eta = 1$, one proximal step lowers the margin from $\gamma_2 = 1$ to 0.012, close to the limit $1/\sqrt{8101} \approx \gamma_2/90$ (panel d). The squared prediction distance (LS) behaves in the same way, and the KL direction error is of exact order $1/t$, as in Theorem 5.3(3) (Figure 6).

7 Conclusion

For **Q1**, when the alignment between the primal and dual steps is bounded below and mild growth and curvature conditions hold, BPPA attains at least the factor c_w of the maximum margin in any norm. For **Q2**, the divergence enters through this factor, which is exact in the worst case for quadratic divergences, and it can also change the type of solution: the KL divergence on the parameters leads to the maximum ℓ_1-margin for every constant stepsize, and, for linearly independent samples and gradient-descent inner steps, the KL divergence on the predictions leads to the minimum-norm label interpolator, which is the maximum ℓ_2-margin classifier if and only if every sample lies on the margin. Open questions include a rate and adaptive stepsizes for Theorem 4.2, adaptive stepsizes under Condition 3, prediction-space KL with more samples than parameters, and a discrete analogue of the horizon-function characterization of mirror flow (Pesme et al., 2024).

AI USE STATEMENT

The authors defined the research problem, identified the relevant lines of prior work, and selected the analytical framework and tools used in the paper. The analysis of strongly convex and smooth divergences builds on proofs developed independently of AI assistance. For other parts of the analysis, the authors developed the arguments iteratively with the assistance of large language models, primarily Claude (Anthropic), with additional assistance from GPT (OpenAI), using them to explore candidate statements and derivations and to check intermediate reasoning. The authors used these outputs as intermediate material in developing, refining, and checking the arguments presented in the paper. The overall structure, theoretical development, and line of argument were determined by the authors. Under their direction, Claude also assisted with drafting the initial manuscript and implementing the numerical experiments, with the resulting material subsequently edited and incorporated by the authors. No datasets were generated with AI tools, no human-subject data are involved, and the authors take full responsibility for the content of the paper.

ETHICS STATEMENT

This is a theoretical study of optimization algorithms for linear classification. It involves no human subjects, personal data or deployed systems, and we do not foresee direct negative societal impact.

REPRODUCIBILITY STATEMENT

All results are stated with explicit assumptions and proved in full in the appendices: Appendix B contains tools from convex analysis, Appendices C and D prove the results of Section 3 for BPPA and mirror descent, and Appendices E and F prove those of Sections 4 and 5. The numerical examples are fully specified in Appendix G, including the datasets, the stepsizes, the numerical solver and the exactly computed reference values; Appendix H describes the CIFAR-100 experiments.

REFERENCES

Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: beyond infinitesimal mirror descent. *International Conference on Machine Learning*, 2021.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Jonathan Eckstein. Nonlinear proximal point algorithms using bregman functions, with applications to convex programming. *Mathematics of Operations Research*, 18(1):202–226, 1993.

Yoav Freund and Robert E Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pp. 1607–1616. PMLR, 2018.

Udaya Ghai, Elad Hazan, and Yoram Singer. Exponentiated gradient meets gradient descent. In *Algorithmic Learning Theory*, 2020.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Sam Green, Craig M Vineyard, and Cetin Kaya Koç. Distillation strategies for proximal policy optimization. *arXiv preprint arXiv:1901.08128*, 2019.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pp. 1832–1841. PMLR, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012.

Daniel Hsu, Vidya Muthukumar, and Ji Xu. On the proliferation of support vectors in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Ziwei Ji and Matus Telgarsky. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pp. 772–804. PMLR, 2021.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2177–2190, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.197. URL https://www.aclweb.org/anthology/2020.acl-main.197.

Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. http://ttic.uchicago.edu/~shai/papers/KakadeShalevTewari09.pdf, 2009. Unpublished manuscript.

Anna Kerekes, Anna Mészáros, and Ferenc Huszár. Depth without the magic: Inductive bias of natural gradient descent. *arXiv preprint arXiv:2111.11542*, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jyrki Kivinen and Manfred K Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

Krzysztof C Kiwiel. Proximal minimization methods with generalized bregman functions. *SIAM journal on control and optimization*, 35(4):1142–1168, 1997.

Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yan Li, Caleb Ju, Ethan X Fang, and Tuo Zhao. Implicit regularization of Bregman proximal point algorithm and mirror descent on separable data. *arXiv preprint arXiv:2108.06808*, 2021.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2018.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 116–131, 2018.

Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: does the loss function matter? *Journal of Machine Learning Research*, 22, 2021.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3420–3428. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/nacson19b.html.

Arkadi S. Nemirovsky and David B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience, 1983.

Scott Pesme, Radu-Alexandru Dragomir, and Nicolas Flammarion. Implicit bias of mirror flow on separable data. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a regularized path to a maximum margin classifier. *The Journal of Machine Learning Research*, 5:941–973, 2004.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, and Nathan Srebro. The implicit bias of gradient descent on separable data. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1q7n9gAb.

Haoyuan Sun, Khashayar Gatmiry, Kwangjun Ahn, and Navid Azizan. A unified approach to controlling implicit regularization via mirror descent. *Journal of Machine Learning Research*, 24 (393):1–58, 2023.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL http://proceedings.mlr.press/v28/sutskever13.html.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017.

Matus Telgarsky. Margins, shrinkage, and boosting. In *International Conference on Machine Learning*, pp. 307–315. PMLR, 2013.

Guanghui Wang, Zihao Hu, Vidya Muthukumar, and Jacob Abernethy. Faster margin maximization rates for generic optimization methods. In *Advances in Neural Information Processing Systems*, volume 36, pp. 62488–62518, 2023.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, 2020.

Jingfeng Wu, Peter L Bartlett, Matus Telgarsky, and Bin Yu. Large stepsize gradient descent for logistic loss: non-monotonicity of the loss improves optimization efficiency. In *Conference on Learning Theory*, 2024.

Huan Xu and Shie Mannor. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

Lei Yang and Kim-Chuan Toh. Bregman proximal point algorithm revisited: a new inexact version and its variant. *arXiv preprint arXiv:2105.10370*, 2021.

Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisit knowledge distillation: a teacher-free framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3649–3658, 2020.

Ruiqi Zhang, Jingfeng Wu, Licong Lin, and Peter L Bartlett. Minimax optimal convergence of gradient descent in logistic regression via large and adaptive stepsizes. *arXiv preprint arXiv:2504.04105*, 2025.

A Summary of Results and Notation

Table 1: Summary of main results. All four methods share the asymptotic margin lower bound $\kappa_w^{-1/2}\gamma^*$ under Condition 2 (Proposition 3.6); the geometries with $c_w = 0$ of Sections 4–5 are summarized in Table 2. The iteration complexity to reach a $(1-\epsilon)$-fraction of this bound, the corresponding bound of Li et al. (2021), and the loss rate are reported. $\widetilde{\mathcal{O}}(\cdot)$ hides polylogarithmic factors; see the proofs for full expressions.

Method	**Iteration to $(1-\epsilon)$-margin**	**Earlier bound**	**Loss rate**	**Ref.**
BPPA, const. η	$\exp(\widetilde{\mathcal{O}}(1/\epsilon))$	$\exp(\widetilde{\mathcal{O}}(1/\epsilon^2))$	$\mathcal{O}(\log^2 t\,/\,t)$	Thm C.1
BPPA, adaptive η_t	$\widetilde{\mathcal{O}}(1/\epsilon^2)$	$\mathcal{O}(1/\epsilon^8)$	$\exp(-\sqrt{t})$	Thm C.2
MD, const. η	$\exp\left(\widetilde{\mathcal{O}}\left(1/\epsilon\right)\right)$	$\exp(\widetilde{\mathcal{O}}(1/\epsilon^{3/2}))$	$\mathcal{O}(\log^2 t\,/\,t)$	Thm D.1
MD, adaptive η_t	$\widetilde{\mathcal{O}}(1/\epsilon^2)$	$\mathcal{O}(1/\epsilon^4)$	$\exp(-\sqrt{t})$	Thm D.2

Notation guide for the appendix

Throughout the proofs we use the following running notation. Fix a reference norm $\|\cdot\|$ and its dual $\|\cdot\|_*$, and let w satisfy Condition 2 with constants μ_w, L_w.

$$\begin{aligned} z_i &:= y_i x_i & \mathcal{L}_t &:= \mathcal{L}(\theta_t) \\ \Delta_t &:= \theta_{t+1} - \theta_t & r_t &:= \mathcal{L}_{t+1}/\mathcal{L}_t \\ \alpha_t &:= \eta_t \mathcal{L}_t & m_T &:= \min_i \langle \theta_T, z_i \rangle \\ Q_T &:= \textstyle\sum_{t<T}(1-r_t)\,\|\Delta_t\| & \kappa &:= L_w/\mu_w \end{aligned}$$

γ^*: maximum $\|\cdot\|_*$-margin; u^*: max-margin direction; $R_* = \max_i \|x_i\|_*$: data radius.

$q(\alpha)$: per-step contraction factor (Theorem C.2); c_t, $\bar{c}_T$, c_w: alignment, its path-weighted average, and the alignment constant (Definition 3.1, Condition 3).

B Tools from Convex Analysis

The first lemma is a folklore result on the duality of Bregman divergence, which shows that the Bregman distance between primal variables (x, y) induced by w is the same as the Bregman distance between the dual variables $(z_y = \nabla w(y), z_x = \nabla w(x))$ induced by w^*.

Lemma B.1 (Duality of Bregman Divergence)**.** *Under Condition 2, let $w^*(\cdot) = \max_x \langle \cdot, x\rangle - w(x)$ be the convex conjugate of w. Then*

$$D_w(x, y) = D_{w^*}(\nabla w(y), \nabla w(x)), \tag{9}$$

where D_w and D_{w^*} denote the Bregman divergence induced by w and w^* respectively.

Proof. By the definition of conjugate function, we have $w^*(v) = \max_u \langle v, u\rangle - w(u)$. Under Condition 2, w is μ_w-strongly convex and L_w-smooth on $\mathbb{R}^d$; strong convexity of w guarantees that

Table 2: Scope of the results. (W): well-posed proximal step; c_w: alignment constant (Condition 3).

DGF	(W)	c_w	margin conclusion
Condition 2	✓	$\geq \kappa_w^{-1/2}$	$\geq c_w \gamma^*$ (Theorems C.1, 3.5)
quadratic $\frac{1}{2}\theta^\top A\theta$, any norm	✓	exact	attained by a one-point dataset (Theorem 3.8); ℓ_2: $\frac{2\sqrt{\kappa}}{1+\kappa}$; (I, ℓ_1): $\frac{2}{1+\sqrt{d}}$
pointwise conditioned Hessian	✓	$\geq \bar{\kappa}^{-1/2}$	$\geq c_w \gamma^*$ (Theorem 3.5); e.g. $\|\theta\|_2^p/p$ (Example 3.4)
separable powers $\frac{1}{p}\|\theta\|_p^p$	✓	0	only $\gamma^* \liminf \bar{c}_T$ (Proposition 3.2); exact ℓ_p known for MD (Sun et al., 2023)
KL / hyperbolic entropy	✓	0	ℓ_1-margin $\to \gamma_1$; limit directions maximize the ℓ_1-margin, with convergence if the maximizer is unique; support coordinates persist (Theorem 4.2); no rate
prediction-space KL, $\mathrm{rank} Z = n \leq d$	×	0	min-norm label interpolator $Z^\dagger \mathbf{1} = X^\dagger y$ under the natural selection; $=$ SVM iff $K^{-1}\mathbf{1} \geq 0$ (Theorem 5.3)
prediction-space LS, $\mathrm{rank} Z = d$	✓	Kantorovich	$\widehat{\Sigma}$-Mahalanobis maximum margin (Corollary 3.7)

the maximizer defining $w^*(v)$ exists and is unique for every v, so w^* is finite and differentiable everywhere, with

$$\nabla w^*(v) = \operatorname*{argmax}_u \langle v, u\rangle - w(u), \Rightarrow \langle v, \nabla w^*(v)\rangle - w(\nabla w^*(v)) = w^*(v), \quad \forall v.$$

Since w is proper, convex, and continuous (all guaranteed by Condition 2), the Fenchel–Moreau theorem gives $w = (w^*)^*$, which also gives

$$\nabla w(u) = \operatorname*{argmax}_v \langle u, v\rangle - w^*(v), \Rightarrow \langle u, \nabla w(u)\rangle - w^*(\nabla w(u)) = w(u),$$
$$\Rightarrow u = \nabla w^*(\nabla w(u)), \quad \forall u.$$

By the definition of D_{w^*}, we have

$$\begin{aligned}
&D_{w^*}(\nabla w(y), \nabla w(x)) \\
&\quad = w^*(\nabla w(y)) - w^*(\nabla w(x)) - \langle \nabla w^*(\nabla w(x)), \nabla w(y) - \nabla w(x)\rangle \\
&\quad = w^*(\nabla w(y)) - w^*(\nabla w(x)) - \langle x, \nabla w(y) - \nabla w(x)\rangle \\
&\quad = w^*(\nabla w(y)) - \langle \nabla w(y), y\rangle - [w^*(\nabla w(x)) - \langle \nabla w(x), x\rangle] - \langle \nabla w(y), x - y\rangle \\
&\quad = w(x) - w(y) - \langle \nabla w(y), x - y\rangle \\
&\quad = D_w(x, y).
\end{aligned}$$

□

The second lemma establishes the duality between smoothness and strong convexity w.r.t. general $\|\cdot\|$-norm. We refer interested readers to Kakade et al. (2009) for the detailed proof.

Lemma B.2 (Theorem 6, Kakade et al. (2009)). *If $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and μ-strongly convex ($\mu > 0$) with respect to $\|\cdot\|$-norm, then $f^* : \mathbb{R}^d \to \mathbb{R}$ is $\frac{1}{\mu}$-smooth and $\frac{1}{L}$ strongly convex with respect to $\|\cdot\|_*$-norm. Here $f^*(\cdot) = \max_x \langle x, \cdot\rangle - f(x)$ denotes the convex conjugate of f, and $(\|\cdot\|, \|\cdot\|_*)$ are a pair of dual norms.*

Lemma B.3 (Primal-dual Bregman coupling). *Under Condition 2, for all $x, y \in \mathbb{R}^d$,*

$$D_w(x, y) \geq \tfrac{1}{2}\sqrt{\mu_w / L_w}\ \|x - y\|\ \|\nabla w(x) - \nabla w(y)\|_* .$$

Proof. By Lemma B.1, $D_w(x, y) = D_{w^*}(\nabla w(y), \nabla w(x))$. Write $D_w(x, y) = \sqrt{D_w(x, y)} \cdot \sqrt{D_{w^*}(\nabla w(y), \nabla w(x))}$. The Bregman sandwich (Condition 2) gives $D_w(x, y) \geq \frac{\mu_w}{2}\|x - y\|^2$; Lemma B.2 gives $D_{w^*} \geq \frac{1}{2L_w}\|\cdot\|_*^2$. Taking square roots and multiplying yields the result. □

The next two lemmas are the ingredients of the alignment bounds of Proposition 3.3.

Lemma B.4 (Co-coercivity of smooth convex functions). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex, differentiable, and L-smooth w.r.t. $\|\cdot\|$ in the sense $D_f(x, y) \le \frac{L}{2}\|x - y\|^2$ for all x, y. Then for all x, y,*

$$D_f(x, y) \;\ge\; \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_*^2, \qquad \langle \nabla f(x) - \nabla f(y), x - y\rangle \;\ge\; \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_*^2.$$

No strict convexity is required.

Proof. Fix y and let $\varphi(z) := f(z) - \langle \nabla f(y), z\rangle$. Then φ is convex, $\nabla\varphi(y) = 0$, so $\varphi(y) = \min\varphi$; moreover $D_\varphi = D_f$ and $\varphi(x) - \varphi(y) = D_f(x, y)$. Let $g := \nabla\varphi(x) = \nabla f(x) - \nabla f(y)$, pick d with $\|d\| = 1$ and $\langle g, d\rangle = \|g\|_*$ (such d exists by the definition of the dual norm), and set $z := x - (\|g\|_*/L)d$. The smoothness bound $D_\varphi(z, x) \le \frac{L}{2}\|z - x\|^2$ gives

$$\varphi(z) \le \varphi(x) + \langle g, z - x\rangle + \frac{L}{2}\|z - x\|^2 = \varphi(x) - \frac{\|g\|_*^2}{L} + \frac{\|g\|_*^2}{2L} = \varphi(x) - \frac{\|g\|_*^2}{2L}.$$

Hence $D_f(x, y) = \varphi(x) - \varphi(y) \ge \varphi(x) - \varphi(z) \ge \|g\|_*^2/(2L)$. Adding the same bound with x and y exchanged and using $D_f(x, y) + D_f(y, x) = \langle \nabla f(x) - \nabla f(y), x - y\rangle$ gives the second claim. □

Lemma B.5 (Kantorovich angle bound). *Let $H \in \mathbb{R}^{d\times d}$ be symmetric positive definite with condition number $\kappa(H) = \lambda_{\max}(H)/\lambda_{\min}(H)$. Then for every $v \ne 0$,*

$$\langle Hv, v\rangle \;\ge\; \frac{2\sqrt{\kappa(H)}}{1 + \kappa(H)}\,\|v\|_2\,\|Hv\|_2\,,$$

with equality for $v \propto \lambda_{\min}^{-1/2} e_{\min} + \lambda_{\max}^{-1/2} e_{\max}$, where $e_{\min}, e_{\max}$ are unit eigenvectors for $\lambda_{\min}, \lambda_{\max}$.

Proof. Put $u = H^{1/2}v$, so that $\langle Hv, v\rangle = \|u\|_2^2$, $\|v\|_2 = \left\|H^{-1/2}u\right\|_2$ and $\|Hv\|_2 = \left\|H^{1/2}u\right\|_2$. Kantorovich's inequality (Horn & Johnson, 2012) states that $\langle Hu, u\rangle \left\langle H^{-1}u, u\right\rangle \le \frac{(\lambda_{\max}+\lambda_{\min})^2}{4\lambda_{\max}\lambda_{\min}}\|u\|_2^4$, i.e. $\left\|H^{1/2}u\right\|_2^2 \left\|H^{-1/2}u\right\|_2^2 \le \frac{(1+\kappa)^2}{4\kappa}\|u\|_2^4$, which rearranges to the claim. Equality in Kantorovich's inequality holds for $u \propto e_{\min} + e_{\max}$, i.e. $v = H^{-1/2}u \propto \lambda_{\min}^{-1/2} e_{\min} + \lambda_{\max}^{-1/2} e_{\max}$. □

C Proofs in Section 3

Reading guide. The BPPA proofs have one central reduction: Proposition C.5 shows that the finite-time margin gap is controlled by two quantities only—a defect term Q_T and the norm growth $\|\theta_T\|$. Theorems C.1 and C.2 therefore split into two independent tasks: bounding Q_T and lower-bounding $\|\theta_T\|$. Constant stepsizes give $Q_T = O(1)$ and $\|\theta_T\| = \Omega(\log T)$; adaptive stepsizes give $Q_T = O(\log T)$ and $\|\theta_T\| = \Omega(\sqrt{T})$. Corollary 3.7 is proved after these two instantiations, followed by the proofs of Proposition 3.2, Proposition 3.3 and Theorem 3.5, which relax Condition 2.

Full statements for strongly convex and smooth DGFs

The two theorems below are the full form of Proposition 3.6(1) and (2). Their asymptotic bound $\kappa_w^{-1/2}\gamma^*$ and the adaptive speed-up are due to Li et al. (2021); the finite-time bounds are sharper than those stated there, $\exp(\widetilde{\mathcal{O}}(1/\epsilon))$ instead of $\exp(\widetilde{\mathcal{O}}(1/\epsilon^2))$ and $\widetilde{\mathcal{O}}(1/\epsilon^2)$ instead of $\mathcal{O}(1/\epsilon^8)$.

Theorem C.1 (Constant stepsize BPPA). *Assume Condition 2, $\theta_0 = 0$ and the exponential loss. For any constant stepsize $\eta_t = \eta > 0$:*

(1) for all $t \ge 1$, $\mathcal{L}(\theta_t) \le \frac{1}{\gamma^*\eta t} + \frac{L_w \log^2(e + \gamma^*\eta t)}{4\gamma^{*2}\eta t} = \mathcal{O}\big(\frac{L_w \log^2(e+\gamma^*\eta t)}{\gamma^{*2}\eta t}\big)$;

(2) the margin satisfies $\min_{i\in[n]} \langle \theta_T/\|\theta_T\|, z_i\rangle \ge \kappa_w^{-1/2}\gamma^* - \mathcal{O}(1/\log(T+1))$; consequently, the number of iterations to reach a $(1-\epsilon)$-fraction of the bound is $T_\epsilon \le \exp(\widetilde{\mathcal{O}}(1/\epsilon))$.

Theorem C.2 (Adaptive stepsize BPPA). *Under the setup of Theorem C.1, let $\eta_t = \alpha_t/\mathcal{L}(\theta_t)$ for a positive sequence $\{\alpha_t\}$.*

(1) For every $t \geq 0$, $\mathcal{L}(\theta_{t+1}) \leq q(\alpha_t)\,\mathcal{L}(\theta_t)$, where $q(\alpha) \coloneqq \min_{u\in(0,1)} \max\left\{u,\ e^{-2\alpha u^2 {\gamma^*}^2/L_w}\right\} < 1$; this part requires only the L_w-smoothness of w.

(2) With $\alpha_t = 1/\sqrt{t+1}$, $\min_i \langle \theta_T/\left\|\theta_T\right\|, z_i\rangle \geq \kappa_w^{-1/2}\gamma^* - \mathcal{O}(\log(T+1)/\sqrt{T+1})$, so $T_\epsilon = \widetilde{\mathcal{O}}(1/\epsilon^2)$, and $\mathcal{L}(\theta_t) = \mathcal{O}(\exp(-{\gamma^*}^2\sqrt{t}/L_w))$.

A few remarks are in order. First, with a constant stepsize the effective step $\eta\mathcal{L}(\theta_t)$ shrinks with the loss and $\|\theta_t\|$ grows only logarithmically, which makes T_ϵ exponential; the adaptive rule $\eta_t = \alpha_t/\mathcal{L}(\theta_t)$ with $\alpha_t = 1/\sqrt{t+1}$ makes $\|\theta_t\|$ grow like $\sqrt{t}$. Second, exact BPPA decreases the loss for every stepsize, so neither theorem needs a cap on η; mirror descent does (Appendix D). Third, the problem does not meet the standard assumptions of the classical BPPA analysis, since no finite minimizer exists; the proof compares the iterates with the scaled classifier $s\,u^*$ for the loss rate and applies Proposition 3.2 for the margin. Fourth, since $q(\alpha) \to 0$ as $\alpha \to \infty$, one proximal step with a large normalized stepsize reduces the loss by an arbitrary factor without instability, the implicit counterpart of large-stepsize gradient descent (Wu et al., 2024; Zhang et al., 2025).

Lemma C.3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex, and $D_w(\theta', \theta) = w(\theta') - w(\theta) - \langle \nabla w(\theta), \theta' - \theta\rangle$ be the Bregman divergence associated with w. Letting*

$$\theta_{t+1} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^d} f(\theta) + \frac{1}{2\eta} D_w(\theta, \theta_t), \tag{10}$$

then for all $\theta \in \mathbb{R}^d$, we have

$$f(\theta_{t+1}) + \frac{1}{2\eta} D_w(\theta, \theta_{t+1}) \leq f(\theta) + \frac{1}{2\eta} D_w(\theta, \theta_t) - \frac{1}{2\eta} D_w(\theta_{t+1}, \theta_t).$$

Proof. Since θ_{t+1} minimizes the sum of convex f and differentiable $(1/2\eta)D_w(\cdot, \theta_t)$ over $\mathbb{R}^d$, the stationarity condition gives: there exists $f'(\theta_{t+1}) \in \partial f(\theta_{t+1})$ such that

$$f'(\theta_{t+1}) + \frac{1}{2\eta}\nabla D_w(\theta_{t+1}, \theta_t) = 0,$$

where $\nabla D_w(\theta_{t+1}, \theta_t)$ denotes the gradient of $D_w(\cdot, \theta_t)$ at θ_{t+1}. By the three-point identity for Bregman divergences:

$$D_w(\theta, \theta_t) = D_w(\theta_{t+1}, \theta_t) + \langle \nabla D_w(\theta_{t+1}, \theta_t), \theta - \theta_{t+1}\rangle + D_w(\theta, \theta_{t+1}).$$

Combining with the convexity of f, which gives $\langle f'(\theta_{t+1}), \theta - \theta_{t+1}\rangle \leq f(\theta) - f(\theta_{t+1})$, we conclude the proof. □

Lemma C.4. *For proximal point algorithm initialized at $\theta_0 = 0$ with Bregman divergence $D_w(\cdot,\cdot)$, where $w(\cdot)$ is L_w-smooth w.r.t. $\|\cdot\|$-norm, we have*

$$\mathcal{L}(\theta_t) - \mathcal{L}(\theta) \leq \frac{L_w}{4\eta t}\left\|\theta\right\|^2.$$

Proof. Given the update rule

$$\theta_{s+1} \in \operatorname*{argmin}_{\theta} \mathcal{L}(\theta) + \frac{1}{2\eta} D_w(\theta, \theta_s),$$

together with Lemma C.3, wherein we take $f(\cdot) = \mathcal{L}(\cdot)$, and $\theta = \theta_s$, we have that

$$\mathcal{L}(\theta_{s+1}) \leq \mathcal{L}(\theta_s) - \frac{1}{2\eta} D_w(\theta_s, \theta_{s+1}) - \frac{1}{2\eta} D_w(\theta_{s+1}, \theta_s). \tag{11}$$

Since $w(\cdot)$ is convex, we have $D_w(\theta_s, \theta_{s+1}) \geq 0$ and $D_w(\theta_{s+1}, \theta_s) \geq 0$. Thus, we have monotone improvement $\mathcal{L}(\theta_{s+1}) \leq \mathcal{L}(\theta_s)$. On the other hand, by Lemma C.3, we also have

$$\mathcal{L}(\theta_{s+1}) - \mathcal{L}(\theta) \leq \frac{1}{2\eta} D_w(\theta, \theta_s) - \frac{1}{2\eta} D_w(\theta, \theta_{s+1}). \tag{12}$$

Summing up the previous inequality from $s = 0$ to $s = t - 1$, we have

$$\sum_{s=1}^{t} \mathcal{L}(\theta_s) - \mathcal{L}(\theta) \leq \frac{1}{2\eta} D_w(\theta, \theta_0).$$

Additionally, since $w(\cdot)$ is L_w-smooth w.r.t. $\|\cdot\|$ norm, we have

$$D_w(\theta, \theta_0) \leq \frac{L_w}{2} \|\theta - \theta_0\|^2 .$$

Combining with the fact $\mathcal{L}(\theta_{s+1}) \leq \mathcal{L}(\theta_s)$ for all $s \geq 0$ and $\theta_0 = 0$, we arrive at

$$\mathcal{L}(\theta_t) - \mathcal{L}(\theta) \leq \frac{L_w}{4\eta t} \|\theta\|^2 .$$

□

Defect decomposition tools

The following tools are used to convert loss decay into finite-time margin bounds for both Theorems C.1 and C.2. Throughout, we write $z_i \coloneqq y_i x_i$ for the signed data points.

By Condition 2, we have the Bregman sandwich: $\frac{\mu_w}{2} \|\theta - \theta'\|^2 \leq D_w(\theta, \theta') \leq \frac{L_w}{2} \|\theta - \theta'\|^2$ for all θ, θ'.

The key idea is to decompose the unnormalized margin $m_T = \min_i \langle \theta_T, z_i \rangle$ into a main term proportional to the path length $\sum_t \|\theta_{t+1} - \theta_t\|$ and a defect term Q_T that measures the cumulative deviation of the loss ratio $r_t = \mathcal{L}(\theta_{t+1})/\mathcal{L}(\theta_t)$ from 1. After normalizing by $\|\theta_T\|$, the margin lower bound reduces to $\sqrt{\mu/L}\,\gamma^*$ minus a remainder controlled by $Q_T/\|\theta_T\|$. Bounding Q_T (via step-length estimates) and $\|\theta_T\|$ (via loss decay) separately then closes the finite-time margin guarantee.

Proposition C.5 (Defect-to-margin reduction)**.** *Let $\{\theta_t\}$ be exact BPPA iterates initialized at $\theta_0 = 0$ with arbitrary positive stepsizes η_t, and define $\alpha_t \coloneqq \eta_t \mathcal{L}(\theta_t)$, $r_t \coloneqq \mathcal{L}(\theta_{t+1})/\mathcal{L}(\theta_t)$, and the absolute defect*

$$Q_T \coloneqq \sum_{t=0}^{T-1} (1 - r_t) \|\theta_{t+1} - \theta_t\| .$$

Then for every $T \geq 1$,

$$\min_{i \in [n]} \left\langle \frac{\theta_T}{\|\theta_T\|}, y_i x_i \right\rangle \geq \sqrt{\frac{\mu_w}{L_w}}\,\gamma^* - \frac{\log(n\mathcal{L}(\theta_0)) + \sqrt{\mu_w/L_w}\,\gamma^*\, Q_T}{\|\theta_T\|}.$$

Proof. **Step 1: Unnormalized margin from loss telescoping.** From Lemma C.3 with $\theta = \theta_t$, the one-step BPPA descent gives $\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t) - \frac{1}{2\eta_t}\big(D_w(\theta_t, \theta_{t+1}) + D_w(\theta_{t+1}, \theta_t)\big)$, hence $\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t) \exp\big(-\frac{D_w(\theta_t,\theta_{t+1}) + D_w(\theta_{t+1},\theta_t)}{2\alpha_t}\big)$ using $1 - x \leq e^{-x}$ and $\alpha_t = \eta_t \mathcal{L}(\theta_t)$. Iterating over $t = 0, \ldots, T-1$:

$$\mathcal{L}(\theta_T) \leq \mathcal{L}(\theta_0) \exp\Big(-\sum_{t=0}^{T-1} \frac{D_w(\theta_t, \theta_{t+1}) + D_w(\theta_{t+1}, \theta_t)}{2\alpha_t}\Big).$$

Since $\mathcal{L}(\theta_T) \geq \frac{1}{n} e^{-\min_i \langle \theta_T, z_i \rangle}$ (where $z_i = y_i x_i$), taking logarithms gives

$$m_T \coloneqq \min_i \langle \theta_T, z_i \rangle \geq -\log(n\mathcal{L}(\theta_0)) + \sum_{t=0}^{T-1} \frac{D_w(\theta_t, \theta_{t+1}) + D_w(\theta_{t+1}, \theta_t)}{2\alpha_t}. \tag{13}$$

Step 2: Bregman-to-displacement bound. By Lemma B.3, $D_w(\theta_t, \theta_{t+1}) \geq \frac{1}{2}\sqrt{\mu_w/L_w}\,\|\theta_{t+1} - \theta_t\|\,\|\nabla w(\theta_{t+1}) - \nabla w(\theta_t)\|_*$. The proximal optimality condition gives $\nabla w(\theta_{t+1}) - \nabla w(\theta_t) = -2\eta_t \nabla \mathcal{L}(\theta_{t+1})$, so $\|\nabla w(\theta_{t+1}) - \nabla w(\theta_t)\|_* = 2\eta_t \|\nabla \mathcal{L}(\theta_{t+1})\|_*$. Since $\|\nabla \mathcal{L}(\theta_{t+1})\|_* \geq \mathcal{L}(\theta_{t+1})\gamma^*$, we obtain

$$D_w(\theta_t, \theta_{t+1}) \geq \alpha_t r_t \gamma^* \sqrt{\frac{\mu_w}{L_w}}\,\|\theta_{t+1} - \theta_t\| .$$

The same bound holds for $D_w(\theta_{t+1}, \theta_t)$ by applying Lemma B.3 with swapped arguments.

Step 3: Combine. Substituting the Step 2 lower bound into (13):

$$m_T \geq -\log(n\mathcal{L}(\theta_0)) + \sqrt{\frac{\mu_w}{L_w}}\,\gamma^* \sum_{t=0}^{T-1} r_t \left\|\theta_{t+1} - \theta_t\right\|.$$

By the definition of Q_T, we split the weighted sum:

$$\sum_{t=0}^{T-1} r_t \left\|\theta_{t+1} - \theta_t\right\| = \sum_{t=0}^{T-1} \left\|\theta_{t+1} - \theta_t\right\| - Q_T \geq \left\|\theta_T\right\| - Q_T,$$

where the last step uses the triangle inequality $\sum_t \|\theta_{t+1} - \theta_t\| \geq \|\theta_T - \theta_0\| = \|\theta_T\|$ (since $\theta_0 = 0$). Dividing both sides by $\|\theta_T\|$ gives the stated bound. □

The next lemma provides the step-length upper bound used to control Q_T in Theorems C.1 and C.2.

Lemma C.6 (Linear step-length bound)**.** *For exact BPPA with arbitrary positive stepsizes η_t,*

$$\left\|\theta_{t+1} - \theta_t\right\| \leq \frac{2\eta_t}{\mu_w} \left\|\nabla\mathcal{L}(\theta_{t+1})\right\|_* \leq \frac{2R_*\eta_t}{\mu_w}\mathcal{L}(\theta_{t+1}).$$

Equivalently, with $\alpha_t = \eta_t\mathcal{L}(\theta_t)$, $\|\theta_{t+1} - \theta_t\| \leq \frac{2R_*}{\mu_w}\alpha_t r_t \leq \frac{2R_*}{\mu_w}\alpha_t$.

Proof. The proximal optimality condition gives $\nabla w(\theta_{t+1}) - \nabla w(\theta_t) = -2\eta_t\nabla\mathcal{L}(\theta_{t+1})$. By μ_w-strong convexity of w, $\langle\nabla w(\theta_{t+1}) - \nabla w(\theta_t), \theta_{t+1} - \theta_t\rangle \geq \mu_w \left\|\theta_{t+1} - \theta_t\right\|^2$; applying $\langle a, b\rangle \leq \|a\|_* \|b\|$ gives $\|\theta_{t+1} - \theta_t\| \leq \frac{1}{\mu_w}\|\nabla w(\theta_{t+1}) - \nabla w(\theta_t)\|_* = \frac{2\eta_t}{\mu_w}\|\nabla\mathcal{L}(\theta_{t+1})\|_*$. For exponential loss, $\|\nabla\mathcal{L}(\theta)\|_* \leq R_*\mathcal{L}(\theta)$. □

The next lemma shows that $1 - r_t$ is controlled by $\|\Delta_t\|$, yielding the bound $Q_T = O(1)$ (constant step) or $Q_T = O(\log T)$ (adaptive step).

Lemma C.7 (Ratio defect bound)**.** *For exponential loss, $1 - r_t \leq R_* \left\|\theta_{t+1} - \theta_t\right\|$, and hence $Q_T \leq R_* \sum_{t=0}^{T-1} \left\|\theta_{t+1} - \theta_t\right\|^2$.*

Proof. Writing $r_t = \sum_i p_{t,i} e^{-\langle\theta_{t+1} - \theta_t, z_i\rangle}$ with weights $p_{t,i} = e^{-\langle\theta_t, z_i\rangle} / \sum_j e^{-\langle\theta_t, z_j\rangle}$, we have $r_t \geq e^{-R_*\|\theta_{t+1} - \theta_t\|}$ since $\langle\theta_{t+1} - \theta_t, z_i\rangle \leq R_* \left\|\theta_{t+1} - \theta_t\right\|$. Therefore $1 - r_t \leq 1 - e^{-R_*\|\theta_{t+1} - \theta_t\|} \leq R_* \left\|\theta_{t+1} - \theta_t\right\|$. □

Proof of Theorem C.1

◇ ***Proof of Theorem C.1-(1).*** Define $u_s = s\,u^*$ for $s > 0$ (recall u^* and γ^* from Definition 2.1). By Lemma C.4, we have $\mathcal{L}(\theta_t) \leq \mathcal{L}(u_s) + \frac{L_w}{4\eta t}\left\|u_s\right\|^2$. On the other hand, we know that

$$\mathcal{L}(u_s) = \frac{1}{n}\sum_{i=1}^{n} \exp\left\{-s\left\langle u^*, y_i x_i\right\rangle\right\} \leq \exp\left(-s\gamma^*\right).$$

Thus, $\mathcal{L}(\theta_t) \leq \exp\left(-s\gamma^*\right) + \frac{s^2 L_w}{4\eta t}$. Finally, by taking $s = \frac{\log(\mathrm{e} + t\gamma^*\eta)}{\gamma^*}$ (where e denotes Euler's number), we have

$$\mathcal{L}(\theta_t) \leq \frac{1}{\gamma^*\eta\,t} + \frac{L_w \log^2\left(e + \gamma^*\eta\,t\right)}{4{\gamma^*}^2\eta\,t} = \mathcal{O}\left\{\frac{L_w \log^2\left(e + \gamma^*\eta\,t\right)}{{\gamma^*}^2\eta\,t}\right\}. \tag{14}$$

□

We proceed to show the parameter convergence of the proximal point algorithm with constant step-size.

◇ ***Proof of Theorem C.1-(2).*** By Proposition C.5, the margin deficit is bounded by

$$\sqrt{\frac{\mu_w}{L_w}}\gamma^* - \min_{i\in[n]} \left\langle \frac{\theta_T}{\|\theta_T\|}, y_i x_i \right\rangle \le \frac{\log(n\mathcal{L}(\theta_0)) + \sqrt{\mu_w/L_w}\,\gamma^*\, Q_T}{\|\theta_T\|}.$$

Therefore, to make this deficit at most $\epsilon\sqrt{\mu_w/L_w}\gamma^*$, it suffices to show that $Q_T = O(1)$ and $\|\theta_T\| \gtrsim \log T$.

Step 1: Bounded defect. By Lemma C.6 with constant stepsize $\eta_t = \eta$, $\|\theta_{t+1} - \theta_t\| \le \frac{2R_*\eta}{\mu_w}\mathcal{L}(\theta_{t+1})$. By Lemma C.7, $Q_T \le R_* \sum_{t=0}^{T-1} \|\theta_{t+1} - \theta_t\|^2 \le \frac{4{R_*}^3\eta^2}{\mu_w^2} \sum_{t=0}^{T-1} \mathcal{L}(\theta_{t+1})^2$. From Part (1), $\mathcal{L}(\theta_t) \le C\log^2(e+t)/(t+1)$, so $\sum_{t=0}^{\infty} \mathcal{L}(\theta_{t+1})^2 \le C' \sum_{t=1}^{\infty} \frac{\log^4(e+t)}{(t+1)^2} < \infty$. Therefore Q_T is uniformly bounded in T.

Step 2: Norm growth. Since $\mathcal{L}(\theta_T) \ge \frac{1}{n} e^{-R_*\|\theta_T\|}$ (because $\langle \theta_T, z_i\rangle \le R_* \|\theta_T\|$ for all i), we obtain

$$\|\theta_T\| \ge \frac{1}{R_*} \log \frac{1}{n\mathcal{L}(\theta_T)}.$$

From Part (1), $\mathcal{L}(\theta_T) = \mathcal{O}(\log^2(e+T)/(T+1))$, hence $\|\theta_T\| = \Omega(\log(T+1))$ with the implicit constant depending on $(R_*, \gamma^*, L_w, \eta, n)$.

Step 3: Combine. By Proposition C.5,

$$\min_{i\in[n]} \left\langle \frac{\theta_T}{\|\theta_T\|}, y_i x_i \right\rangle \ge \sqrt{\frac{\mu_w}{L_w}}\gamma^* - \frac{C_0}{\log(T+1)},$$

where C_0 depends on $(R_*, \gamma^*, L_w, \mu_w, \eta, n)$. Solving $C_0/\log(T+1) \le \epsilon$ gives $T_\epsilon \le \exp(\mathcal{O}(1/\epsilon))$. □

PROOF OF THEOREM C.2

◇ ***Proof of Theorem C.2-(1).*** By (11), we have $\mathcal{L}(\theta_{s+1}) \le \mathcal{L}(\theta_s) - \frac{1}{2\eta_s} D_w(\theta_s, \theta_{s+1}) - \frac{1}{2\eta_s} D_w(\theta_{s+1}, \theta_s)$. Recall that our stepsize is given by $\eta_s = \frac{\alpha_s}{\mathcal{L}(\theta_s)}$, which gives us

$$\begin{aligned}
\mathcal{L}(\theta_{s+1}) &\le \mathcal{L}(\theta_s) - \frac{1}{2\eta_s} D_w(\theta_s, \theta_{s+1}) - \frac{1}{2\eta_s} D_w(\theta_{s+1}, \theta_s) \qquad (15)\\
&\le \mathcal{L}(\theta_s) \exp\left\{ -\frac{1}{2\eta_s\mathcal{L}(\theta_s)} D_w(\theta_s, \theta_{s+1}) - \frac{1}{2\eta_s\mathcal{L}(\theta_s)} D_w(\theta_{s+1}, \theta_s) \right\}\\
&= \mathcal{L}(\theta_s) \exp\left\{ -\frac{1}{2\alpha_s} D_w(\theta_s, \theta_{s+1}) - \frac{1}{2\alpha_s} D_w(\theta_{s+1}, \theta_s) \right\}. \qquad (16)
\end{aligned}$$

By Lemma B.2 (w^* is $(1/L_w)$-strongly convex w.r.t. $\|\cdot\|_*$),

$$\begin{aligned}
D_w(\theta_s, \theta_{s+1}) &= D_{w^*}(\nabla w(\theta_{s+1}), \nabla w(\theta_s))\\
&\ge \frac{1}{2L_w} \|\nabla w(\theta_{s+1}) - \nabla w(\theta_s)\|_*^2\\
&\ge \frac{2\eta_s^2}{L_w} \mathcal{L}^2(\theta_{s+1}) {\gamma^*}^2 \qquad (17)\\
&= \frac{2\alpha_s^2}{L_w} \left(\frac{\mathcal{L}(\theta_{s+1})}{\mathcal{L}(\theta_s)}\right)^2 {\gamma^*}^2,
\end{aligned}$$

where the first equality holds by Lemma B.1, and the second inequality holds by the dual-norm bound $\|\nabla\mathcal{L}(\theta)\|_* \ge \mathcal{L}(\theta)\gamma^*$. By the same argument applied to $D_w(\theta_{s+1}, \theta_s) = D_{w^*}(\nabla w(\theta_s), \nabla w(\theta_{s+1}))$ (Lemma B.1 and the $(1/L_w)$-strong convexity of w^*), we also have $D_w(\theta_{s+1}, \theta_s) \ge \frac{2\alpha_s^2}{L_w} \left(\frac{\mathcal{L}(\theta_{s+1})}{\mathcal{L}(\theta_s)}\right)^2 {\gamma^*}^2$. Hence we obtain

$$\frac{\mathcal{L}(\theta_{s+1})}{\mathcal{L}(\theta_s)} \le \exp\left\{ -\left(\frac{\mathcal{L}(\theta_{s+1})}{\mathcal{L}(\theta_s)}\right)^2 \frac{2\alpha_s{\gamma^*}^2}{L_w} \right\}.$$

We conclude that $r_s := \mathcal{L}(\theta_{s+1})/\mathcal{L}(\theta_s) < 1$. Now for any $u \in (0,1)$, if $r_s \geq u$, then

$$r_s = \frac{\mathcal{L}(\theta_{s+1})}{\mathcal{L}(\theta_s)} \leq \exp\left(-u^2 \frac{2\alpha_s \gamma^{*2}}{L_w}\right).$$

Hence we have

$$\frac{\mathcal{L}(\theta_{s+1})}{\mathcal{L}(\theta_s)} = r_s \leq \max\left\{u, \exp\left(-u^2 \frac{2\alpha_s \gamma^{*2}}{L_w}\right)\right\}, \qquad \forall u \in (0,1).$$

Since the previous inequality holds for any $u \in (0,1)$, we minimize the right hand side with respect to u, and obtain

$$\frac{\mathcal{L}(\theta_{s+1})}{\mathcal{L}(\theta_s)} \leq \min_{u\in(0,1)} \max\left\{u, \exp\left(-u^2 \frac{2\alpha_s \gamma^{*2}}{L_w}\right)\right\} := q(\alpha_s). \tag{18}$$

Hence loss $\{\mathcal{L}(\theta_s)\}_{s\geq 0}$ contracts with contraction coefficient $q(\alpha_s)$. □

◇ ***Loss-decay part of Theorem C.2-(2).*** To show loss convergence, by (18), we have

$$\mathcal{L}(\theta_t) \leq \exp\left\{\sum_{s=0}^{t-1} \log q(\alpha_s)\right\}.$$

Now by the definition of $q(\alpha_s)$, taking $u = \frac{1}{2}$ in the min-max gives $q(\alpha_s) \leq \max\{\frac{1}{2}, \exp(-\frac{\alpha_s \gamma^{*2}}{2L_w})\}$. For all s such that $\alpha_s \leq \frac{2L_w \log 2}{\gamma^{*2}}$ (which holds for $s \geq s_0 := \lceil \frac{\gamma^{*4}}{4L_w^2 \log^2 2} \rceil$), the exponential term exceeds $\frac{1}{2}$, so $q(\alpha_s) \leq \exp(-\frac{\alpha_s \gamma^{*2}}{2L_w})$. Multiplying over $s = s_0, \ldots, t-1$ and using $q(\alpha_s) < 1$ for $s < s_0$, we get

$$\mathcal{L}(\theta_t) \leq \mathcal{L}(\theta_0) \prod_{s=0}^{t-1} q(\alpha_s) \leq C \exp\left(-\sum_{s=s_0}^{t-1} \frac{\alpha_s \gamma^{*2}}{2L_w}\right) = \mathcal{O}\left(\exp\left(-\frac{\gamma^{*2}}{L_w}\sqrt{t}\right)\right).$$

□

◇ ***Margin part of Theorem C.2-(2).*** By Proposition C.5, the margin deficit is bounded by $(\log(n\mathcal{L}(\theta_0)) + \sqrt{\mu/L}\,\gamma^* Q_T)/\|\theta_T\|$. With the adaptive schedule $\alpha_t = (t+1)^{-1/2}$, the defect Q_T grows only logarithmically while $\|\theta_T\|$ grows as $\sqrt{T}$ (from the stretched-exponential loss decay), so the ratio vanishes. Specifically, it suffices to show that $Q_T = O(\log T)$ and $\|\theta_T\| = \Omega(\sqrt{T})$.

Step 1: Logarithmic defect. By Lemma C.6 with $\alpha_t = (t+1)^{-1/2}$, $\|\theta_{t+1} - \theta_t\| \leq \frac{2R_*}{\mu_w}\alpha_t$. By Lemma C.7, $Q_T \leq R_* \sum_{t=0}^{T-1} \|\theta_{t+1} - \theta_t\|^2 \leq \frac{4{R_*}^3}{\mu_w^2} \sum_{t=0}^{T-1} \alpha_t^2 = \frac{4{R_*}^3}{\mu_w^2} \sum_{t=0}^{T-1} \frac{1}{t+1} \lesssim \log(T+1)$.

Step 2: Norm growth. The loss-decay part proved above gives $\mathcal{L}(\theta_T) \lesssim e^{-c\sqrt{T+1}}$. Since $\mathcal{L}(\theta_T) \geq \frac{1}{n} e^{-R_*\|\theta_T\|}$ by the definition of exponential loss, we obtain $\|\theta_T\| \geq \frac{1}{R_*} \log \frac{1}{n\mathcal{L}(\theta_T)} \geq \frac{c}{R_*}\sqrt{T+1} - O(1) \gtrsim \sqrt{T+1}$.

Step 3: Combine. By Proposition C.5,

$$\min_{i\in[n]} \left\langle \frac{\theta_T}{\|\theta_T\|}, y_i x_i \right\rangle \geq \sqrt{\frac{\mu_w}{L_w}}\gamma^* - O\left(\frac{\log(T+1)}{\sqrt{T+1}}\right).$$

Solving $\log(T+1)/\sqrt{T+1} \leq \epsilon$ gives $T_\epsilon = \widetilde{\mathcal{O}}(1/\epsilon^2)$. □

PROOF OF COROLLARY 3.7

Proof. Define $\phi(u) = \min_{i\in[n]} \langle u, y_i x_i \rangle$. We observe that $\phi(\cdot)$ is concave, and $\phi(\cdot)$ is positively homogeneous.

By the previous observations, we claim two properties of $\mathcal{U}^* = \operatorname{argmax}_{\|u\|\leq 1} \phi(u)$:

(a) $\mathcal{U}^*$ is singleton, i.e., $\mathcal{U}^* = \{u^*\}$.

(b) $\|u^*\| = 1$.

To show (a), suppose that there exist $u_1, u_2 \in \mathcal{U}^*$ and $u_1 \neq u_2$. Let $u_\alpha = \alpha u_1 + (1-\alpha)u_2$ for any $\alpha \in (0,1)$. Since $\|\cdot\|$ is a strictly convex norm, the unit ball $\{u : \|u\| \leq 1\}$ is a strictly convex set. Hence $u_\alpha \in \operatorname{int}(\{u : \|u\| \leq 1\})$ and $\|u_\alpha\| < 1$. In addition $\phi(u_\alpha) \geq \alpha\phi(u_1) + (1-\alpha)\phi(u_2) = \phi(u_1)$ given the fact that $\phi(\cdot)$ is concave and $\phi(u_1) = \phi(u_2)$.

Then we have

$$\phi\left(\frac{u_\alpha}{\|u_\alpha\|}\right) = \frac{1}{\|u_\alpha\|}\phi(u_\alpha) \geq \frac{1}{\|u_\alpha\|}\phi(u_1) > \phi(u_1),$$

which contradicts the fact that $u_1 \in \mathcal{U}^*$. Thus (a) holds.

On the other hand, using positive homogeneity, for any $u_* \in \mathcal{U}^*$ such that $\|u_*\| < 1$, $\overline{u}_* = u_*/\|u_*\|$ satisfies $\phi(\overline{u}_*) > \phi(u_*)$ and $\|\overline{u}_*\| = 1$, which again yields a contradiction. Thus (b) holds.

To show asymptotic convergence, Theorem C.1-(2) gives $\liminf_{t\to\infty} \phi\big(\frac{\theta_t}{\|\theta_t\|}\big) \geq \gamma^*$. Since $\phi(u) \leq \gamma^*$ for any $\|u\| \leq 1$ by the definition of γ^*, we obtain $\lim_{t\to\infty} \phi\big(\frac{\theta_t}{\|\theta_t\|}\big) = \gamma^*$. Given that $\mathcal{U}^* = \{u^*\}$ is a singleton and $\phi(\cdot)$ is continuous, we must have $\lim_{t\to\infty} \frac{\theta_t}{\|\theta_t\|} = u^*$. □

PROOF OF PROPOSITION 3.2 (MASTER INEQUALITY)

Proof. Fix t and write $\Delta_t = \theta_{t+1} - \theta_t$, $\Delta_t^* = \nabla w(\theta_{t+1}) - \nabla w(\theta_t) = -2\eta_t \nabla\mathcal{L}(\theta_{t+1})$ (optimality condition (4); the minimizer exists and is unique by assumption).

Step 1: loss telescoping. By convexity of $\mathcal{L}$ at θ_{t+1}, $\mathcal{L}(\theta_t) \geq \mathcal{L}(\theta_{t+1}) + \langle \nabla\mathcal{L}(\theta_{t+1}), \theta_t - \theta_{t+1}\rangle = \mathcal{L}(\theta_{t+1}) + \frac{1}{2\eta_t}\langle \Delta_t^*, \Delta_t\rangle$, where we substituted $\nabla\mathcal{L}(\theta_{t+1}) = -\Delta_t^*/(2\eta_t)$. Since $\langle \Delta_t^*, \Delta_t\rangle = D_w(\theta_{t+1}, \theta_t) + D_w(\theta_t, \theta_{t+1}) \geq 0$ by convexity of w, we obtain, with $\alpha_t = \eta_t\mathcal{L}(\theta_t)$ and $1 - x \leq e^{-x}$,

$$\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t)\Big(1 - \frac{\langle \Delta_t^*, \Delta_t\rangle}{2\alpha_t}\Big) \leq \mathcal{L}(\theta_t)\exp\Big(-\frac{\langle \Delta_t^*, \Delta_t\rangle}{2\alpha_t}\Big).$$

Iterating over $t < T$ and using $\mathcal{L}(\theta_T) \geq \frac{1}{n}e^{-\min_i\langle \theta_T, z_i\rangle}$ (Lemma C.8(b) below) gives

$$\min_{i\in[n]}\langle \theta_T, z_i\rangle \;\geq\; -\log\big(n\mathcal{L}(\theta_0)\big) + \sum_{t<T}\frac{\langle \Delta_t^*, \Delta_t\rangle}{2\alpha_t}. \tag{19}$$

Step 2: alignment. By Definition 3.1, the optimality condition and the dual-norm bound $\|\nabla\mathcal{L}(\theta)\|_* \geq \gamma^*\mathcal{L}(\theta)$ (Lemma C.8(a)),

$$\begin{aligned}\langle \Delta_t^*, \Delta_t\rangle &= c_t\|\Delta_t\|\,\|\Delta_t^*\|_* = c_t\|\Delta_t\| \cdot 2\eta_t\|\nabla\mathcal{L}(\theta_{t+1})\|_* \\ &\geq 2\eta_t\gamma^* c_t\|\Delta_t\|\,\mathcal{L}(\theta_{t+1}) = 2\alpha_t\gamma^* c_t r_t\|\Delta_t\|,\end{aligned}$$

where $r_t = \mathcal{L}(\theta_{t+1})/\mathcal{L}(\theta_t)$. Dividing by $2\alpha_t$ and inserting into (19) gives the first form of (5) (with $\mathcal{L}(\theta_0) = 1$ for $\theta_0 = 0$). The second form follows from $\sum_{t<T} r_t\|\Delta_t\| = \sum_{t<T}\|\Delta_t\| - \sum_{t<T}(1-r_t)\|\Delta_t\| = P_T - Q_T$ and the definition of $\bar{c}_T$.

Step 3: normalization. Since $\theta_0 = 0$, the triangle inequality gives $\|\theta_T\| \leq P_T$, and $0 \leq Q_T \leq P_T$ since $0 < r_t \leq 1$; dividing (5) by $\|\theta_T\|$ yields $\min_i\langle \theta_T/\|\theta_T\|, z_i\rangle \geq \gamma^*\bar{c}_T(P_T - Q_T)/P_T - \log n/\|\theta_T\|$.

Step 4: the defect. By Lemma C.8(c), $r_t \geq e^{-R_*\|\Delta_t\|}$, so $1 - r_t \leq 1 - e^{-R_*\|\Delta_t\|} \leq R_*\|\Delta_t\|$ and $Q_T \leq R_*\sum_{t<T}\|\Delta_t\|^2$. If $\mathcal{L}(\theta_t) \to 0$ then $\|\theta_T\| \to \infty$ by Lemma C.8(b), hence $P_T \geq \|\theta_T\| \to \infty$. If moreover $\|\Delta_t\| \to 0$, then for every $\varepsilon > 0$ there is T_0 with $\|\Delta_t\| \leq \varepsilon$ for $t \geq T_0$, and $\sum_{t<T}\|\Delta_t\|^2 \leq \sum_{t<T_0}\|\Delta_t\|^2 + \varepsilon P_T$, so $\limsup_T Q_T/P_T \leq R_*\varepsilon$; letting $\varepsilon \to 0$ gives $Q_T/P_T \to 0$ and (6). □

Lemma C.8 (Elementary facts used above)**.** *For all $\theta, \Delta \in \mathbb{R}^d$: (a) $\gamma^*\mathcal{L}(\theta) \leq \|\nabla\mathcal{L}(\theta)\|_* \leq R_*\mathcal{L}(\theta)$; (b) $\frac{1}{n}e^{-m(\theta)} \leq \mathcal{L}(\theta) \leq e^{-m(\theta)}$ with $m(\theta) = \min_i\langle \theta, z_i\rangle$, hence $\|\theta\| \geq \frac{1}{R_*}\log\frac{1}{n\mathcal{L}(\theta)}$; (c) $\mathcal{L}(\theta + \Delta) \geq e^{-R_*\|\Delta\|}\mathcal{L}(\theta)$.*

Proof. (a) $\nabla\mathcal{L}(\theta) = -\frac{1}{n}\sum_i e^{-\langle\theta,z_i\rangle} z_i$, so $\|\nabla\mathcal{L}(\theta)\|_* \geq \langle -\nabla\mathcal{L}(\theta), u^*\rangle = \frac{1}{n}\sum_i e^{-\langle\theta,z_i\rangle}\langle z_i, u^*\rangle \geq \gamma^*\mathcal{L}(\theta)$ because $\|u^*\| = 1$ and $\langle z_i, u^*\rangle \geq \gamma^*$; the upper bound is the triangle inequality with $\|z_i\|_* \leq R_*$. (b) A sum of n positive terms is at least its largest term and at most n times it; and $m(\theta) \leq \langle\theta, z_1\rangle \leq R_*\|\theta\|$. (c) Termwise, $e^{-\langle\theta+\Delta,z_i\rangle} = e^{-\langle\theta,z_i\rangle}e^{-\langle\Delta,z_i\rangle} \geq e^{-\langle\theta,z_i\rangle}e^{-R_*\|\Delta\|}$. □

Remark C.9. Under Condition 2, Lemma B.3 gives $c_t \geq \sqrt{\mu_w/L_w}$ for every t, and Proposition 3.2 reduces to Proposition C.5; the proof above shows that the conjugate duality of Lemmas B.1–B.2 is only needed to bound the alignment, not for the margin inequality itself.

Proof of Proposition 3.3

Proof. Write $\Delta = x - y$ and $\Delta^* = \nabla w(x) - \nabla w(y)$ for $x \neq y$, and recall $\langle\Delta^*, \Delta\rangle = D_w(x,y) + D_w(y,x)$.

(i) Strong convexity gives $D_w(x,y) \geq \frac{\mu_w}{2}\|\Delta\|^2$ and $D_w(y,x) \geq \frac{\mu_w}{2}\|\Delta\|^2$, so $\langle\Delta^*,\Delta\rangle \geq \mu_w\|\Delta\|^2$. Smoothness and Lemma B.4 give $\langle\Delta^*,\Delta\rangle \geq \frac{1}{L_w}\|\Delta^*\|_*^2$. Multiplying the two lower bounds and taking square roots, $\langle\Delta^*,\Delta\rangle \geq \sqrt{\mu_w\|\Delta\|^2 \cdot L_w^{-1}\|\Delta^*\|_*^2} = \sqrt{\mu_w/L_w}\,\|\Delta\|\,\|\Delta^*\|_*$, i.e. $c(x,y) \geq \sqrt{\mu_w/L_w}$ for every pair.

(ii) Let N be a Borel null set outside which $\lambda > 0$. The ratio defining c_w is undefined for pairs with $\nabla w(x) = \nabla w(y)$, and we exclude such pairs from the infimum; they never occur as consecutive BPPA iterates, since $\Delta_t^* = -2\eta_t\nabla\mathcal{L}(\theta_{t+1})$ and $\|\nabla\mathcal{L}(\theta)\|_* \geq \gamma^*\mathcal{L}(\theta) > 0$ (Lemma C.8(a)). The argument below uses only two regularity properties: ∇w is continuous, and on every segment $s \mapsto \nabla w(y+s\Delta)$ is absolutely continuous with derivative $\nabla^2 w(y+s\Delta)\Delta$ for almost every s. Both hold for $w \in C^2$.

Step 1: segments that meet N in a null set. Fix $x \neq y$, write $\Delta = x - y$, and suppose that $\{s \in [0,1] : y + s\Delta \in N\}$ has one-dimensional measure zero. Then $\Delta^* = \int_0^1 \nabla^2 w(y+s\Delta)\Delta\, ds = \bar{H}\Delta$ with $\bar{H} := \int_0^1 \nabla^2 w(y+s\Delta)\, ds$, and $\bar{\lambda} := \int_0^1 \lambda(y+s\Delta)\, ds > 0$ because the integrand is positive for almost every s. Integrating the pointwise sandwich $\lambda(\theta)\|h\|^2 \leq \langle\nabla^2 w(\theta)h, h\rangle \leq \bar{\kappa}\lambda(\theta)\|h\|^2$ along the segment, for every h,

$$\bar{\lambda}\|h\|^2 \leq \langle\bar{H}h, h\rangle \leq \bar{\kappa}\bar{\lambda}\|h\|^2.$$

(The same scalar function λ appears on both sides, so the ratio $\bar{\kappa}$ survives the averaging even though the levels $\lambda(\theta)$ vary along the segment.) The quadratic function $q(h) := \frac{1}{2}\langle\bar{H}h, h\rangle$ therefore satisfies Condition 2 with constants $(\bar{\lambda}, \bar{\kappa}\bar{\lambda})$, and $\nabla q(\Delta) - \nabla q(0) = \bar{H}\Delta = \Delta^*$. Applying part (i) to q at the pair $(\Delta, 0)$ gives $\langle\Delta^*,\Delta\rangle \geq \bar{\kappa}^{-1/2}\|\Delta\|\,\|\Delta^*\|_*$, i.e. $c(x,y) \geq \bar{\kappa}^{-1/2}$. If $\|\cdot\| = \|\cdot\|_2$, then $\bar{H}$ is symmetric positive definite with all eigenvalues in $[\bar{\lambda}, \bar{\kappa}\bar{\lambda}]$, so its condition number is at most $\bar{\kappa}$, and Lemma B.5 gives $\langle\bar{H}\Delta,\Delta\rangle \geq \frac{2\sqrt{\kappa(\bar{H})}}{1+\kappa(\bar{H})}\|\Delta\|_2\|\bar{H}\Delta\|_2 \geq \frac{2\sqrt{\bar{\kappa}}}{1+\bar{\kappa}}\|\Delta\|_2\|\bar{H}\Delta\|_2$, because $\kappa \mapsto 2\sqrt{\kappa}/(1+\kappa)$ is decreasing on $[1,\infty)$.

Step 2: all pairs. A null set N can contain an entire segment, so Step 1 does not cover every pair directly; we reach the remaining pairs by translation. Fix $x \neq y$ with $\nabla w(x) \neq \nabla w(y)$ and $\Delta = x - y$. The set $\mathcal{E} := \{(\epsilon, s) \in \mathbb{R}^d \times [0,1] : y + \epsilon + s\Delta \in N\}$ is Borel, and for each fixed s its ϵ-section is the null set $N - y - s\Delta$. By Fubini's theorem $\mathcal{E}$ is null, so for almost every ϵ the s-section $\{s : y + \epsilon + s\Delta \in N\}$ is null. For every such ϵ, Step 1 applies to the translated pair $(x+\epsilon, y+\epsilon)$, which has the same Δ. Choose such $\epsilon_k \to 0$. By continuity of ∇w, $\Delta_k^* := \nabla w(x+\epsilon_k) - \nabla w(y+\epsilon_k) \to \Delta^* \neq 0$, so $c(x+\epsilon_k, y+\epsilon_k) \to c(x,y)$, and the bounds of Step 1 pass to the limit. Hence $c_w \geq \bar{\kappa}^{-1/2}$, and $c_w \geq 2\sqrt{\bar{\kappa}}/(1+\bar{\kappa})$ in the Euclidean case.

Radial powers. For $w(\theta) = \frac{1}{p}\|\theta\|_2^p$ and $\theta \neq 0$, $\nabla^2 w(\theta) = \|\theta\|_2^{p-2}\left(I + (p-2)\widehat{\theta}\widehat{\theta}^\top\right)$ with $\widehat{\theta} = \theta/\|\theta\|_2$; its eigenvalues are $\|\theta\|_2^{p-2}$ (multiplicity $d-1$) and $(p-1)\|\theta\|_2^{p-2}$. Hence the sandwich of (ii) holds with $\lambda(\theta) = \min(1, p-1)\|\theta\|_2^{p-2}$ and $\bar{\kappa} = \max(p-1, \frac{1}{p-1})$, with $N = \{0\}$. Every segment meets N in at most one point, so Step 1 applies directly. For $p \geq 2$, $w \in C^2$. For $1 < p < 2$, w is not C^2 at the origin, but $\nabla w(\theta) = \|\theta\|_2^{p-2}\theta$ is continuous, and on a segment through the origin at $s = s_0$ the operator norm $\|\nabla^2 w(y+s\Delta)\| = \|y + s\Delta\|_2^{p-2} = \|\Delta\|_2^{p-2}|s - s_0|^{p-2}$ is integrable

in s because $p-2>-1$; hence $s\mapsto \nabla w(y+s\Delta)$ is absolutely continuous and the two regularity properties used in the proof hold. Therefore $c_w \ge 2\sqrt{\bar\kappa}/(1+\bar\kappa) = 2\sqrt{p-1}/p$ in both cases. □

Example C.10 (Two families with $c_w = 0$)**.** *(a)* Separable powers $w(\theta) = \frac{1}{p}\sum_i |\theta_i|^p$, $p>2$, *with* $\|\cdot\| = \|\cdot\|_2$. *Take* $x=(\varepsilon, M)$ *and* $y=(-\varepsilon, M+\varepsilon^{p-1})$ *in* $\mathbb{R}^2$. *Then* $\Delta = (2\varepsilon, -\varepsilon^{p-1})$ *and* $\Delta^* = (2\varepsilon^{p-1},\ M^{p-1}-(M+\varepsilon^{p-1})^{p-1}) \approx (2\varepsilon^{p-1}, -(p-1)M^{p-2}\varepsilon^{p-1})$, *so*

$$c(x,y) \approx \frac{4\varepsilon^p + (p-1)M^{p-2}\varepsilon^{2p-2}}{2\varepsilon\cdot(p-1)M^{p-2}\varepsilon^{p-1}} = \frac{2}{(p-1)M^{p-2}} + \frac{\varepsilon^{p-2}}{2} \longrightarrow 0$$

as $M\to\infty$ *and then* $\varepsilon\to 0$. *The mechanism is that the curvature* $|\theta_i|^{p-2}$ *of one coordinate can be made arbitrarily larger than that of another. (b)* Hyperbolic entropy w_β *of Section 4, with* $\|\cdot\| = \|\cdot\|_2$ *and* $\beta_i\equiv\beta$. *Take* $x=(M,0)$, $y=(M+1,\varepsilon)$ *with* $\varepsilon = M^{-1/2}$. *Then* $\Delta=(1,\varepsilon)$ *and* $\Delta^* = (\operatorname{arcsinh}\frac{M+1}{2\beta} - \operatorname{arcsinh}\frac{M}{2\beta},\ \operatorname{arcsinh}\frac{\varepsilon}{2\beta}) \approx (1/M,\ \varepsilon/2\beta)$, *so* $c(x,y)\approx \frac{1/M+\varepsilon^2/2\beta}{\sqrt{1+\varepsilon^2}\sqrt{1/M^2+\varepsilon^2/4\beta^2}} \approx \frac{1/M+\varepsilon^2/2\beta}{\varepsilon/2\beta} = \frac{2\beta}{M\varepsilon}+\varepsilon\to 0$. *By norm equivalence (Section 3),* $c_w=0$ *for every norm in both cases.*

Proof of Theorem 3.5

We first record two lemmas that replace the roles played by global strong convexity in the proof of Theorem C.1: an a priori upper bound on $\|\theta_t\|$ (which under Condition 2 follows from $D_w(su^*,\theta_t)\ge\frac{\mu_w}{2}\|su^*-\theta_t\|^2$) and a bound on the step length (Lemma C.6 under Condition 2).

Lemma C.11 (A priori norm bound)**.** *Assume the proximal step is well posed,* $\theta_0=0$, *constant stepsize* η, *and let* $G(s):=D_w(su^*,0)$. *For* $t\ge 1$ *and* $s_t:=\log t/\gamma^*$,

$$D_w(s_tu^*,\theta_t) \ \le\ B_t := G(s_t)+2\eta. \tag{20}$$

Consequently, writing $N_t=\|\theta_t\|$: *(a) under local strong convexity with decay,* $D_w(x,y)\ge\frac{\lambda_0}{2}(1+\max(\|x\|,\|y\|))^{-a}\|x-y\|^2$ *with* $a<2$, *one has* $N_t\le\max\{2s_t,\ 1,\ (2^{a+3}B_t/\lambda_0)^{1/(2-a)}\}$; *(b) under uniform convexity* $D_w(x,y)\ge\frac{\mu_p}{p}\|x-y\|^p$, *one has* $N_t\le s_t+(pB_t/\mu_p)^{1/p}$. *In both cases* $N_t=\mathcal{O}(\operatorname{polylog} t)$ *when* $G(s)\le C_G(1+s)^\rho$.

Proof. Summing the three-point inequality (12) with $\theta=s_tu^*$ over $k=0,\dots,t-1$ and dropping the non-negative terms $\mathcal{L}(\theta_{k+1})$, $D_w(s_tu^*,\theta_t)\le D_w(s_tu^*,\theta_0)+2\eta\sum_{k<t}\big[\mathcal{L}(s_tu^*)-\mathcal{L}(\theta_{k+1})\big]\le G(s_t)+2\eta t\,\mathcal{L}(s_tu^*)\le G(s_t)+2\eta t\,e^{-s_t\gamma^*}=G(s_t)+2\eta$, using $\mathcal{L}(su^*)=\frac{1}{n}\sum_i e^{-s\langle u^*,z_i\rangle}\le e^{-s\gamma^*}$. This is (20). (a) If $N_t\le\max(2s_t,1)$ there is nothing to prove. Otherwise $N_t-s_t\ge N_t/2$ and $1+\max(N_t,s_t)=1+N_t\le 2N_t$, so the decay condition and (20) give $\frac{\lambda_0}{2}(2N_t)^{-a}(N_t/2)^2\le B_t$, i.e. $N_t^{2-a}\le 2^{a+3}B_t/\lambda_0$. (b) The uniform convexity bound gives $\frac{\mu_p}{p}(N_t-s_t)_+^p\le\frac{\mu_p}{p}\|s_tu^*-\theta_t\|^p\le B_t$ (recall $\|u^*\|=1$, so $\|s_tu^*-\theta_t\|\ge N_t-s_t$). □

The decay condition in (a) follows from $\langle\nabla^2 w(\theta)h,h\rangle\ge\lambda_0(1+\|\theta\|)^{-a}\|h\|^2$ by integrating along the segment $[y,x]$, on which $\|y+s(x-y)\|\le\max(\|x\|,\|y\|)$ by convexity of the norm; it also implies the strong monotonicity $\langle\nabla w(x)-\nabla w(y),x-y\rangle\ge\lambda_0(1+\max(\|x\|,\|y\|))^{-a}\|x-y\|^2$ used next.

Lemma C.12 (Vanishing steps)**.** *Under the hypotheses of Lemma C.11: (a) under local strong convexity with decay,* $\|\Delta_t\|\le\frac{2\eta R_*}{\lambda_0}\big(1+\max(N_t,N_{t+1})\big)^a\mathcal{L}(\theta_{t+1})$; *(b) under uniform convexity of degree* p, $\|\Delta_t\|\le\big(\frac{p\eta R_*}{\mu_p}\mathcal{L}(\theta_{t+1})\big)^{1/(p-1)}$. *Combined with the loss rate of Theorem 3.5(1) and Lemma C.11: in case (a),* $\|\Delta_t\|=\mathcal{O}(\operatorname{polylog}(t)/t)$ *and* $\sum_t\|\Delta_t\|^2<\infty$; *in case (b),* $\|\Delta_t\|\to 0$, *and* $\sum_t\|\Delta_t\|^2<\infty$ *when* $p<3$.

Proof. By the optimality condition and Lemma C.8(a), $\langle\Delta_t^*,\Delta_t\rangle\le\|\Delta_t^*\|_*\|\Delta_t\|=2\eta\|\nabla\mathcal{L}(\theta_{t+1})\|_*\|\Delta_t\|\le 2\eta R_*\mathcal{L}(\theta_{t+1})\|\Delta_t\|$. In case (a), strong monotonicity gives $\langle\Delta_t^*,\Delta_t\rangle\ge\lambda_0(1+\max(N_t,N_{t+1}))^{-a}\|\Delta_t\|^2$; in case (b), $\langle\Delta_t^*,\Delta_t\rangle=D_w(\theta_{t+1},\theta_t)+D_w(\theta_t,\theta_{t+1})\ge\frac{2\mu_p}{p}\|\Delta_t\|^p$. Comparing the two bounds gives the claims. For the summability statements,

$\mathcal{L}(\theta_{t+1}) = \mathcal{O}(\log^\rho t/t)$ by Theorem 3.5(1) and $N_t = \mathcal{O}(\mathrm{polylog}\, t)$ by Lemma C.11; in case (b), $\|\Delta_t\|^2 = \mathcal{O}((\log^\rho t/t)^{2/(p-1)})$ is summable iff $2/(p-1) > 1$, i.e. $p < 3$. □

Proof of Theorem 3.5. (1) By Lemma C.4 (whose proof uses only the three-point inequality and monotone descent, with $\frac{L_w}{4\eta t}\|\theta\|^2$ replaced by $\frac{1}{2\eta t} D_w(\theta, \theta_0)$), $\mathcal{L}(\theta_t) \le \mathcal{L}(su^*) + \frac{G(s)}{2\eta t} \le e^{-s\gamma^*} + \frac{C_G(1+s)^\rho}{2\eta t}$ for every $s \ge 0$; take $s = \log(2\eta t)/\gamma^*$. (2) Condition 3 gives $c_t \ge c_w$ for every t, hence $\bar{c}_T \ge c_w$. By (1), $\mathcal{L}(\theta_t) \to 0$; by Lemma C.12, $\|\Delta_t\| \to 0$. Proposition 3.2, (6), gives $\liminf_T \min_i \langle \theta_T/\|\theta_T\|, z_i\rangle \ge \gamma^* \liminf_T \bar{c}_T \ge c_w\gamma^*$. (3) If $S_\infty := \sum_t \|\Delta_t\|^2 < \infty$, then $Q_T \le R_* S_\infty$ and $P_T \ge \|\theta_T\| \ge \frac{1}{R_*}\log\frac{1}{n\mathcal{L}(\theta_T)}$ (Lemma C.8(b)), so Proposition 3.2 gives

$$\min_i \langle \theta_T/\|\theta_T\|, z_i\rangle \;\ge\; c_w\gamma^* - \frac{c_w\gamma^* {R_*}^2 S_\infty + R_* \log n}{\log\frac{1}{n\mathcal{L}(\theta_T)}},$$

and $\log\frac{1}{n\mathcal{L}(\theta_T)} = \log T - \mathcal{O}(\log\log T)$ by (1). Lemma C.12 shows that $S_\infty < \infty$ in the stated cases. Finally, under Condition 2, Proposition 3.3(i) gives $c_w \ge \sqrt{\mu_w/L_w}$, the growth condition holds with $G(s) \le \frac{L_w}{2}s^2\|u^*\|^2 = \frac{L_w}{2}s^2$, and the decay condition holds with $\lambda_0 = \mu_w$, $a = 0$. □

Proof of Theorem 3.8

Proof. For $w(\theta) = \frac{1}{2}\theta^\top A\theta$ one has $\nabla w(x) - \nabla w(y) = A(x-y)$, so $c(x,y)$ depends only on $\Delta = x - y$ and $c_w = \inf_{\Delta\neq 0}\langle A\Delta, \Delta\rangle/(\|\Delta\|\,\|A\Delta\|_*)$. The ratio is invariant under scaling of Δ and is a continuous positive function on the compact unit sphere $\{\|\Delta\| = 1\}$, so the infimum is a minimum, attained at some Δ° with $\|\Delta^\circ\| = 1$.

Let $z = A\Delta^\circ$ and consider the one-point dataset $\{(z, +1)\}$, so that $\mathcal{L}(\theta) = e^{-\langle\theta, z\rangle}$ (the two-point dataset $\{(z,+1),(-z,-1)\}$ gives the same loss since $y_i x_i = z$ for both points). The optimality condition (4) is $A(\theta_{t+1} - \theta_t) = 2\eta e^{-\langle\theta_{t+1}, z\rangle} z$, i.e. $\theta_{t+1} - \theta_t = 2\eta e^{-\langle\theta_{t+1}, z\rangle} A^{-1}z = 2\eta e^{-\langle\theta_{t+1}, z\rangle}\Delta^\circ$. Starting from $\theta_0 = 0$, induction gives $\theta_t = \rho_t\Delta^\circ$ with $\rho_0 = 0$ and $\rho_{t+1} = \rho_t + 2\eta\exp(-\rho_{t+1}\langle\Delta^\circ, z\rangle)$; the increments are positive, so ρ_t increases, and if ρ_t stayed bounded its increments would be bounded below by a positive constant, a contradiction; hence $\rho_t \uparrow \infty$. Now for every $t \ge 1$,

$$\min_{(x,y)\in\mathcal{S}} \langle \theta_t/\|\theta_t\|, yx\rangle = \frac{\langle\Delta^\circ, z\rangle}{\|\Delta^\circ\|} = \langle A\Delta^\circ, \Delta^\circ\rangle, \qquad \gamma^* = \max_{\|u\|\le 1}\langle u, z\rangle = \|z\|_* = \|A\Delta^\circ\|_*,$$

so the margin ratio equals $\langle A\Delta^\circ, \Delta^\circ\rangle/(\|\Delta^\circ\|\,\|A\Delta^\circ\|_*) = c(\Delta^\circ, 0) = c_w$, exactly and at every iteration.

Euclidean norm. For $\|\cdot\| = \|\cdot\|_2$, $c_w = \min_{\|\Delta\|_2=1}\langle A\Delta, \Delta\rangle / \|A\Delta\|_2 = \frac{2\sqrt{\kappa}}{1+\kappa}$ by Lemma B.5 (the bound and its equality case), with $\Delta^\circ \propto \lambda_{\min}^{-1/2} e_{\min} + \lambda_{\max}^{-1/2} e_{\max}$.

$A = I$ and $\|\cdot\| = \|\cdot\|_1$ on $\mathbb{R}^d$. Here $c_w = \min_{\Delta\neq 0}\|\Delta\|_2^2/(\|\Delta\|_1\,\|\Delta\|_\infty)$. By scaling and symmetry we may assume $\Delta_1 = \|\Delta\|_\infty = 1$ and $\Delta_i \ge 0$; let $S = \sum_{i\ge 2}\Delta_i \in [0, d-1]$. By the Cauchy–Schwarz inequality, $\sum_{i\ge 2}\Delta_i^2 \ge S^2/(d-1)$ with equality when all Δ_i ($i \ge 2$) equal $S/(d-1) \le 1$. Hence $c_w = \min_{S\in[0,d-1]}\frac{1+S^2/(d-1)}{1+S}$; the derivative of $S \mapsto \frac{1+S^2/(d-1)}{1+S}$ vanishes when $S^2 + 2S - (d-1) = 0$, i.e. at $S = \sqrt{d} - 1$, where the value is $\frac{1+(\sqrt{d}-1)^2/(d-1)}{\sqrt{d}} = \frac{1+(\sqrt{d}-1)/(\sqrt{d}+1)}{\sqrt{d}} = \frac{2}{1+\sqrt{d}}$, and it is a minimum since the ratio equals 1 at $S = 0$ and $\frac{d}{d} = 1$ at $S = d-1$. The minimizer is $\Delta^\circ = (1, \frac{1}{1+\sqrt{d}}, \ldots, \frac{1}{1+\sqrt{d}})$. □

D Mirror Descent: Full Statements and Proofs

This appendix states the mirror-descent results summarized in Proposition 3.6(3) and Section 3.4 in full and proves them. MD (3) has the optimality condition $\nabla w(\theta_{t+1}) - \nabla w(\theta_t) = -2\eta_t\nabla\mathcal{L}(\theta_t)$, with the gradient at the current iterate.

Theorem D.1 (Constant Stepsize MD). *Assume Condition 2, $\theta_0 = 0$ and the exponential loss. For any constant stepsize $\eta_t = \eta \le \mu_w/(2{R_*}^2)$:*

(1) $\lim_{t\to\infty} \mathcal{L}(\theta_t) = 0$, and for all $t \ge 1$, $\mathcal{L}(\theta_t) \le \frac{1}{\gamma^* \eta t} + \frac{L_w \log^2(e+\gamma^* \eta t)}{4{\gamma^*}^2 \eta t}$;

(2) $\min_i \langle \theta_T/\|\theta_T\|, z_i\rangle \ge \kappa_w^{-1/2}\gamma^* - \mathcal{O}(1/\log(T+1))$, so $T_\epsilon \le \exp(\widetilde{\mathcal{O}}(1/\epsilon))$.

Theorem D.2 (Varying Stepsize MD). *Under the conditions of Theorem D.1, let $\eta_t = \alpha_t/\mathcal{L}(\theta_t)$ with $\alpha_t = \min\{\mu_w/(2{R_*}^2), 1/\sqrt{t+1}\}$. Then*

(1) $\mathcal{L}(\theta_t) = \mathcal{O}(\exp(-{\gamma^*}^2\sqrt{t}/L_w))$;

(2) $\min_i \langle \theta_T/\|\theta_T\|, z_i\rangle \ge \kappa_w^{-1/2}\gamma^* - \mathcal{O}(\log(T+1)/\sqrt{T+1})$, so $T_\epsilon = \widetilde{\mathcal{O}}(1/\epsilon^2)$.

Corollary D.3 (Quadratic/Mahalanobis DGF for MD). *Let $\|\cdot\| = \|\cdot\|_A$ for some $A \succ 0$. Under the conditions of Theorem D.1, MD with $w(\theta) = \langle \theta, A\theta\rangle$ converges to the maximum $\|\cdot\|_{A^{-1}}$-margin classifier, $\lim_t \min_i \langle \theta_t/\|\theta_t\|, z_i\rangle = \gamma^*$ and $\lim_t \theta_t/\|\theta_t\| = u^*$. In this case MD is steepest descent in the A-metric, recovering Gunasekar et al. (2018).*

Proposition 3.6(3) consists of Theorems D.1 and D.2; Remark D.6 at the end of this appendix shows that $\kappa_w^{-1/2}$ may be replaced by c_w in Theorem D.1, and Corollary D.3 and Proposition D.5 give the quadratic case.

Reading guide. The mirror-descent analysis differs from BPPA in only one place: the linearized update introduces a stability term that requires a small-step condition for monotone descent (Theorem D.1-(1), Step 1). After that, the margin analysis follows the same normalized-path template as BPPA, with the defect Q_T replaced by a weighted defect ratio E_T/P_T where $P_T = \sum_{t<T}\|\Delta_t\|$ and $E_T = \sum_{t<T}\frac{\mathcal{L}(\theta_t)\eta_t {R_*}^2}{\mu_w}\|\Delta_t\|$. Theorems D.1 and D.2 are obtained by showing $E_T/P_T \to 0$.

Proof of Theorem D.1

We prove parts (1) and (2) separately. Part (1) establishes monotone descent and the $\mathcal{O}(\log^2 t/t)$ loss rate via a comparison argument; Part (2) derives the margin lower bound using the normalized-path template.

$\diamond$ *Proof of Theorem D.1-(1):* We first show that for small enough stepsize η_t, we have monotonic improvement, i.e., $\mathcal{L}(\theta_{t+1}) \le \mathcal{L}(\theta_t)$. Suppose not, and we have $\mathcal{L}(\theta_{t+1}) > \mathcal{L}(\theta_t)$, then given the second-order taylor expansion, there exists $\widetilde{\theta}_t = \tau_t\theta_t + (1-\tau_t)\theta_{t+1}$ for some $\tau_t \in [0,1]$, such that

$$\mathcal{L}(\theta_{t+1}) \le \mathcal{L}(\theta_t) + \langle \nabla\mathcal{L}(\theta_t), \theta_{t+1}-\theta_t\rangle + \frac{1}{2}(\theta_{t+1}-\theta_t)^\top \nabla^2\mathcal{L}(\widetilde{\theta}_t)(\theta_{t+1}-\theta_t).$$

By some calculation, we have

$$\nabla^2\mathcal{L}(\widetilde{\theta}_t) = \frac{1}{n}\sum_{i=1}^n \exp\left(-\left\langle \widetilde{\theta}_t, y_i x_i\right\rangle\right) x_i x_i^\top.$$

Since $h^\top x_i x_i^\top h = \langle x_i, h\rangle^2 \le \|x_i\|_*^2\|h\|^2 \le {R_*}^2\|h\|^2$, we obtain $h^\top\nabla^2\mathcal{L}(\widetilde{\theta}_t)h \le {R_*}^2\mathcal{L}(\widetilde{\theta}_t)\|h\|^2$. Since $\mathcal{L}(\cdot)$ is a convex function, by the definition of $\widetilde{\theta}_t$ we have

$$\mathcal{L}(\widetilde{\theta}_t) \le \tau_t\mathcal{L}(\theta_t) + (1-\tau_t)\mathcal{L}(\theta_{t+1}) \le \max\{\mathcal{L}(\theta_t), \mathcal{L}(\theta_{t+1})\} = \mathcal{L}(\theta_{t+1}),$$

where the last inequality holds by our assumption $\mathcal{L}(\theta_{t+1}) > \mathcal{L}(\theta_t)$. Hence we have

$$\begin{aligned}\mathcal{L}(\theta_{t+1}) &\le \mathcal{L}(\theta_t) + \langle \nabla\mathcal{L}(\theta_t), \theta_{t+1}-\theta_t\rangle + \frac{\max\{\mathcal{L}(\theta_{t+1}), \mathcal{L}(\theta_t)\}}{2}{R_*}^2\|\theta_{t+1}-\theta_t\|^2 \qquad (21)\\ &= \mathcal{L}(\theta_t) + \langle \nabla\mathcal{L}(\theta_t), \theta_{t+1}-\theta_t\rangle + \frac{\mathcal{L}(\theta_{t+1})}{2}{R_*}^2\|\theta_{t+1}-\theta_t\|^2.\end{aligned}$$

To proceed, we employ the well known three-point lemma in mirror descent analysis, which states the following.

Lemma D.4. *Given Bregman divergence $D_w(\cdot,\cdot)$ and mirror descent update*

$$\theta_{t+1} = \operatorname*{argmin}_{\theta} \langle \nabla \mathcal{L}(\theta_t), \theta - \theta_t \rangle + \frac{1}{2\eta_t} D_w(\theta, \theta_t),$$

we have

$$\langle \nabla \mathcal{L}(\theta_t), \theta_{t+1} - \theta \rangle + \frac{1}{2\eta_t} D_w(\theta_{t+1}, \theta_t) \le \frac{1}{2\eta_t} D_w(\theta, \theta_t) - \frac{1}{2\eta_t} D_w(\theta, \theta_{t+1}), \quad \forall \theta.$$

Proof. This is a special case of Lemma C.3 with $f(\theta) = \langle \nabla \mathcal{L}(\theta_t), \theta - \theta_t \rangle$. □

We now establish monotone descent and the comparison recursion.

Step 1: One-step descent (induction on loss monotonicity). We prove $\mathcal{L}(\theta_{t+1}) \le \mathcal{L}(\theta_t)$ for all $t \ge 0$ by induction. The base case is vacuous. Suppose $\mathcal{L}(\theta_s) \le \mathcal{L}(\theta_0)$ for all $s \le t$. If $\mathcal{L}(\theta_{t+1}) \le \mathcal{L}(\theta_t)$, descent holds. Otherwise $\mathcal{L}(\theta_{t+1}) > \mathcal{L}(\theta_t)$, and by (21) (the Taylor expansion with $\max\{\mathcal{L}(\theta_{t+1}), \mathcal{L}(\theta_t)\} = \mathcal{L}(\theta_{t+1})$ in the Hessian coefficient):

$$\begin{aligned}\mathcal{L}(\theta_{t+1}) &\le \mathcal{L}(\theta_t) + \langle \nabla \mathcal{L}(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{\mathcal{L}(\theta_{t+1})}{2} {R_*}^2 \|\theta_{t+1} - \theta_t\|^2 \\ &\le \mathcal{L}(\theta_t) + \langle \nabla \mathcal{L}(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{\mathcal{L}(\theta_{t+1})}{\mu_w} {R_*}^2 D_w(\theta_{t+1}, \theta_t), \end{aligned} \tag{22}$$

where the second inequality uses $D_w(\theta_{t+1}, \theta_t) \ge \frac{\mu_w}{2} \|\theta_{t+1} - \theta_t\|^2$ (Condition 2). Taking $\theta = \theta_t$ in Lemma D.4 gives $\langle \nabla \mathcal{L}(\theta_t), \theta_{t+1} - \theta_t \rangle \le -\frac{1}{2\eta}\big(D_w(\theta_{t+1}, \theta_t) + D_w(\theta_t, \theta_{t+1})\big)$. Substituting into (22) and rearranging:

$$\mathcal{L}(\theta_{t+1})\Big(1 - \frac{{R_*}^2}{\mu_w} D_w(\theta_{t+1}, \theta_t)\Big) \le \mathcal{L}(\theta_t) - \frac{1}{2\eta}\big(D_w(\theta_{t+1}, \theta_t) + D_w(\theta_t, \theta_{t+1})\big). \tag{23}$$

We verify two conditions under $\eta \le \frac{\mu_w}{2{R_*}^2}$:

- *Condition (i):* $\frac{{R_*}^2}{\mu_w} D_w(\theta_{t+1}, \theta_t) < 1$. From the MD optimality, $\nabla w(\theta_{t+1}) - \nabla w(\theta_t) = -2\eta \nabla \mathcal{L}(\theta_t)$; since w^* is $(1/\mu_w)$-smooth w.r.t. $\|\cdot\|_*$, $D_w(\theta_{t+1}, \theta_t) = D_{w^*}(\nabla w(\theta_t), \nabla w(\theta_{t+1})) \le \frac{1}{2\mu_w} \|\nabla w(\theta_{t+1}) - \nabla w(\theta_t)\|_*^2 = \frac{2\eta^2}{\mu_w} \|\nabla \mathcal{L}(\theta_t)\|_*^2 \le \frac{2\eta^2 {R_*}^2}{\mu_w} \mathcal{L}^2(\theta_t)$. By the inductive hypothesis $\mathcal{L}(\theta_t) \le \mathcal{L}(\theta_0) = 1$ and the step-size bound $\eta \le \frac{\mu_w}{2{R_*}^2}$: $\frac{{R_*}^2}{\mu_w} D_w(\theta_{t+1}, \theta_t) \le \frac{2\eta^2 {R_*}^4}{\mu_w^2} \le \frac{1}{2} < 1$.

- *Condition (ii):* $\frac{{R_*}^2 \mathcal{L}(\theta_t)}{\mu_w} \le \frac{1}{2\eta}$. This is equivalent to $\eta \le \frac{\mu_w}{2\mathcal{L}(\theta_t){R_*}^2}$, which holds since $\mathcal{L}(\theta_t) \le \mathcal{L}(\theta_0) = 1$ and $\eta \le \frac{\mu_w}{2{R_*}^2}$. This gives $\frac{{R_*}^2 \mathcal{L}(\theta_t)}{\mu_w} D_w(\theta_{t+1}, \theta_t) \le \frac{1}{2\eta} D_w(\theta_{t+1}, \theta_t) \le \frac{1}{2\eta}\big(D_w(\theta_{t+1}, \theta_t) + D_w(\theta_t, \theta_{t+1})\big)$.

Condition (i) ensures the factor $(1 - \frac{{R_*}^2}{\mu_w} D_w(\theta_{t+1}, \theta_t))$ in (23) is positive; Condition (ii) ensures the RHS of (23) is at most $\mathcal{L}(\theta_t)(1 - \frac{{R_*}^2}{\mu_w} D_w(\theta_{t+1}, \theta_t))$. Dividing both sides of (23) by this positive factor gives $\mathcal{L}(\theta_{t+1}) \le \mathcal{L}(\theta_t)$, contradicting the assumption. This closes the induction.

Step 2: Comparison recursion. With descent established, we have $\mathcal{L}(\theta_{t+1}) \le \mathcal{L}(\theta_t)$ for all t. Now the Hessian bound (21) simplifies to $\max\{\mathcal{L}(\theta_{t+1}), \mathcal{L}(\theta_t)\} = \mathcal{L}(\theta_t)$, so

$$\begin{aligned}\mathcal{L}(\theta_{t+1}) &\le \mathcal{L}(\theta_t) + \langle \nabla \mathcal{L}(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{\mathcal{L}(\theta_t)}{\mu_w} {R_*}^2 D_w(\theta_{t+1}, \theta_t) \\ &\le \mathcal{L}(\theta_t) + \langle \nabla \mathcal{L}(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{1}{2\eta} D_w(\theta_{t+1}, \theta_t), \end{aligned} \tag{24}$$

where the last step uses Condition (ii). By Lemma D.4, for any θ,

$$\langle \nabla \mathcal{L}(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{1}{2\eta} D_w(\theta_{t+1}, \theta_t) \le \langle \nabla \mathcal{L}(\theta_t), \theta - \theta_t \rangle + \frac{1}{2\eta} D_w(\theta, \theta_t) - \frac{1}{2\eta} D_w(\theta, \theta_{t+1}).$$

Together with the convexity bound $\mathcal{L}(\theta_t) + \langle \nabla\mathcal{L}(\theta_t), \theta - \theta_t\rangle \le \mathcal{L}(\theta)$, we obtain

$$\mathcal{L}(\theta_{t+1}) \le \mathcal{L}(\theta) + \frac{1}{2\eta} D_w(\theta, \theta_t) - \frac{1}{2\eta} D_w(\theta, \theta_{t+1}),$$

which is the same recursion (12) as in Lemma C.4. Following the exact same lines as the proof of Theorem C.1-(1), we obtain

$$\mathcal{L}(\theta_t) \le \frac{1}{\gamma^* \eta\, t} + \frac{L_w \log^2(e + \gamma^* \eta\, t)}{4\gamma^{*2} \eta\, t} = \mathcal{O}\left\{ \frac{L_w \log^2(e + \gamma^* \eta\, t)}{\gamma^{*2} \eta\, t} \right\}. \tag{25}$$

□

◇ *Proof of Theorem D.1-(2):* By the proof of Theorem D.1-(1), whenever $\eta = \eta_t \le \frac{\mu_w}{2R_*^2}$, we have $\mathcal{L}(\theta_{t+1}) \le \mathcal{L}(\theta_t)$ for all $t \ge 0$. With descent established, $\max\{\mathcal{L}(\theta_{t+1}), \mathcal{L}(\theta_t)\} = \mathcal{L}(\theta_t)$, so (21) gives

$$\mathcal{L}(\theta_{t+1}) \le \mathcal{L}(\theta_t) + \langle \nabla\mathcal{L}(\theta_t), \theta_{t+1} - \theta_t\rangle + \frac{\mathcal{L}(\theta_t)}{\mu_w} R_*^2 D_w(\theta_{t+1}, \theta_t).$$

Taking $\theta = \theta_t$ in Lemma D.4: $\langle \nabla\mathcal{L}(\theta_t), \theta_{t+1} - \theta_t\rangle \le -\frac{1}{2\eta}\big(D_w(\theta_{t+1}, \theta_t) + D_w(\theta_t, \theta_{t+1})\big)$. Substituting and rearranging:

$$\begin{aligned}
\mathcal{L}(\theta_{t+1}) &\le \mathcal{L}(\theta_t) - \frac{1}{2\eta} D_w(\theta_t, \theta_{t+1}) - \left(\frac{1}{2\eta} - \frac{\mathcal{L}(\theta_t)}{\mu_w} R_*^2\right) D_w(\theta_{t+1}, \theta_t) \qquad (26)\\
&\le \mathcal{L}(\theta_t) \exp\left\{ -\frac{1}{2\eta\mathcal{L}(\theta_t)} D_w(\theta_t, \theta_{t+1}) - \frac{1}{\mathcal{L}(\theta_t)} \left(\frac{1}{2\eta} - \frac{\mathcal{L}(\theta_t)}{\mu_w} R_*^2\right) D_w(\theta_{t+1}, \theta_t) \right\}\\
&\le \mathcal{L}(\theta_0) \exp\left\{ -\sum_{s=0}^{t} \frac{1}{2\eta\mathcal{L}(\theta_s)} D_w(\theta_s, \theta_{s+1}) - \sum_{s=0}^{t} \frac{1}{\mathcal{L}(\theta_s)} \left(\frac{1}{2\eta} - \frac{\mathcal{L}(\theta_s)}{\mu_w} R_*^2\right) D_w(\theta_{s+1}, \theta_s) \right\} \qquad (27)
\end{aligned}$$

By Lemma B.3 and the MD optimality condition $\nabla w(\theta_{s+1}) - \nabla w(\theta_s) = -2\eta\nabla\mathcal{L}(\theta_s)$:

$$\begin{aligned}
D_w(\theta_s, \theta_{s+1}) &\ge \tfrac{1}{2}\sqrt{\mu_w/L_w}\ \|\theta_{s+1} - \theta_s\|\ \|\nabla w(\theta_{s+1}) - \nabla w(\theta_s)\|_*\\
&= \eta\sqrt{\mu_w/L_w}\ \|\theta_{s+1} - \theta_s\|\ \|\nabla\mathcal{L}(\theta_s)\|_*\\
&\ge \eta\sqrt{\mu_w/L_w}\ \mathcal{L}(\theta_s)\gamma^*\ \|\theta_{s+1} - \theta_s\|,
\end{aligned}$$

where the last step uses $\|\nabla\mathcal{L}(\theta_s)\|_* \ge \mathcal{L}(\theta_s)\gamma^*$. The same bound holds for $D_w(\theta_{s+1}, \theta_s)$ by applying Lemma B.3 with swapped arguments. Substituting both Bregman lower bounds into (27), we obtain

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \exp\left(-\langle \theta_{t+1}, y_i x_i\rangle\right)\\
&\le \mathcal{L}(\theta_0) \exp\left\{ -\sum_{s=0}^{t} \frac{1}{2\eta\mathcal{L}(\theta_s)} D_w(\theta_s, \theta_{s+1}) - \sum_{s=0}^{t} \frac{1}{\mathcal{L}(\theta_s)} \left(\frac{1}{2\eta} - \frac{\mathcal{L}(\theta_s)}{\mu_w} R_*^2\right) D_w(\theta_{s+1}, \theta_s) \right\}\\
&\le \mathcal{L}(\theta_0) \exp\left\{ -\sum_{s=0}^{t} \frac{1}{2}\sqrt{\frac{\mu_w}{L_w}}\ \|\theta_{s+1} - \theta_s\|\,\gamma^* \right.\\
&\qquad \left. -\sum_{s=0}^{t} \left(\frac{1}{2} - \frac{\mathcal{L}(\theta_s)}{\mu_w}\eta R_*^2\right) \sqrt{\frac{\mu_w}{L_w}}\gamma^*\ \|\theta_{s+1} - \theta_s\| \right\}.
\end{aligned}$$

Taking logarithm and dividing by $\|\theta_{t+1}\|$ on both sides, we have for any $i \in [n]$,

$$\begin{aligned}
&\left\langle \frac{\theta_{t+1}}{\|\theta_{t+1}\|}, y_i x_i \right\rangle\\
&\ge \frac{-\log n - \log\mathcal{L}(\theta_0)}{\|\theta_{t+1}\|} + \sqrt{\frac{\mu_w}{L_w}}\gamma^* \frac{\sum_{s=0}^{t} \|\theta_{s+1} - \theta_s\|}{\|\theta_{t+1}\|}
\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{\|\theta_{t+1}\|} \sum_{s=0}^{t} \frac{\mathcal{L}(\theta_s)}{\mu_w} \eta {R_*}^2 \sqrt{\frac{\mu_w}{L_w}} \gamma^* \|\theta_{s+1} - \theta_s\| \\
&= \frac{-\log n - \log \mathcal{L}(\theta_0)}{\|\theta_{t+1}\|} + \sqrt{\frac{\mu_w}{L_w}} \gamma^* \frac{\sum_{s=0}^{t} \left(1 - \frac{\mathcal{L}(\theta_s) \eta {R_*}^2}{\mu_w}\right) \|\theta_{s+1} - \theta_s\|}{\|\theta_{t+1}\|} \\
&\geq \sqrt{\frac{\mu_w}{L_w}} \gamma^* + \underbrace{\frac{-\log n - \log \mathcal{L}(\theta_0)}{\|\theta_{t+1}\|}}_{\text{(I)}} - \underbrace{\sqrt{\frac{\mu_w}{L_w}} \gamma^* \frac{\sum_{s=0}^{t} \frac{\mathcal{L}(\theta_s) \eta {R_*}^2}{\mu_w} \|\theta_{s+1} - \theta_s\|}{\sum_{s=0}^{t} \|\theta_{s+1} - \theta_s\|}}_{\text{(II)}}.
\end{aligned} \tag{28}
$$

To obtain the last inequality, write $S_t = \sum_{s=0}^{t} \|\theta_{s+1} - \theta_s\|$ and $A_t = \sum_{s=0}^{t} \frac{\mathcal{L}(\theta_s) \eta {R_*}^2}{\mu_w} \|\theta_{s+1} - \theta_s\|$. Since the stability condition gives $\frac{\mathcal{L}(\theta_s) \eta {R_*}^2}{\mu_w} \leq \frac{1}{2}$ for all s, we have $S_t - A_t \geq 0$. Therefore $\frac{S_t - A_t}{\|\theta_{t+1}\|} \geq \frac{S_t - A_t}{S_t} = 1 - \frac{A_t}{S_t}$, where we used $\|\theta_{t+1}\| \leq S_t$ (triangle inequality).

To get asymptotic convergence, observe that by (25), $\mathcal{L}(\theta_t) \to 0$. Thus $\|\theta_{t+1}\| \to \infty$ and the error weight $\mathcal{L}(\theta_s) \eta {R_*}^2 / \mu_w \to 0$, so (I)$\to 0$ and (II)$\to 0$, giving

$$
\liminf_{t \to \infty} \left\langle \frac{\theta_{t+1}}{\|\theta_{t+1}\|}, y_i x_i \right\rangle \geq \sqrt{\frac{\mu_w}{L_w}} \gamma^*.
$$

To get the finite-time margin gap, we bound (I) and (II) separately.

Bounding (I). From the loss convergence (25), $\mathcal{L}(\theta_t) = \mathcal{O}(\log^2 t / t)$. Since $\mathcal{L}(\theta_t) \geq \frac{1}{n} e^{-R_* \|\theta_t\|}$, we obtain

$$
\|\theta_{t+1}\| \geq \frac{1}{R_*} \log \frac{1}{n \mathcal{L}(\theta_{t+1})} \geq \frac{1}{R_*} \log \frac{1}{n \mathcal{L}(\theta_t)} = \Omega(\log t), \tag{29}
$$

Therefore (I) $= \mathcal{O}(1/\log t)$.

Bounding (II). By the MD optimality and strong monotonicity of ∇w: $\|\theta_{s+1} - \theta_s\| \leq \frac{2\eta R_*}{\mu_w} \mathcal{L}(\theta_s)$. Hence the numerator of (II) satisfies $\sum_{s \leq t} \frac{\mathcal{L}(\theta_s) \eta {R_*}^2}{\mu_w} \|\Delta_s\| \leq \frac{2\eta^2 {R_*}^3}{\mu_w^2} \sum_{s \leq t} \mathcal{L}(\theta_s)^2 \leq \frac{2\eta^2 {R_*}^3}{\mu_w^2} \sum_{s \geq 0} \mathcal{L}(\theta_s)^2 < \infty$, where the last sum converges by (25), while the denominator is $\sum_{s \leq t} \|\Delta_s\| \geq \|\theta_{t+1}\| = \Omega(\log t)$ by (29). Therefore (II) $= \mathcal{O}(1/\log t)$.

Combining: the margin gap is $\mathcal{O}(1/\log(T+1))$. Solving for the gap to drop below ϵ gives $T_\epsilon \leq \exp(\widetilde{\mathcal{O}}(1/\epsilon))$. □

Proof of Theorem D.2

We prove part (1) first, as the loss decay it establishes is needed for the norm-growth bound in part (2).

◇ *Proof of Theorem D.2-(1):* By the stepsize choice $\alpha_s = \min\{\frac{\mu_w}{2{R_*}^2}, \frac{1}{\sqrt{s+1}}\}$, the stability condition $\eta_s \leq \mu_w / (2\mathcal{L}(\theta_s) {R_*}^2)$ of Theorem D.1 is satisfied at every step, so the one-step descent and comparison recursion from that proof apply. Iterating the exponential form (as in the proof of Theorem D.1-(2)):

$$
\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_0) \exp\Big(-\sum_{s=0}^{t} \frac{1}{2\alpha_s} D_w(\theta_s, \theta_{s+1})\Big).
$$

We lower bound $D_w(\theta_s, \theta_{s+1})$. By Lemma B.1, $D_w(\theta_s, \theta_{s+1}) = D_{w^*}(\nabla w(\theta_{s+1}), \nabla w(\theta_s)) \geq \frac{1}{2L_w} \|\nabla w(\theta_{s+1}) - \nabla w(\theta_s)\|_*^2$. The MD optimality gives $\nabla w(\theta_{s+1}) - \nabla w(\theta_s) = -2\eta_s \nabla \mathcal{L}(\theta_s)$, and $\|\nabla \mathcal{L}(\theta_s)\|_* \geq \mathcal{L}(\theta_s) \gamma^*$, so

$$
D_w(\theta_s, \theta_{s+1}) \geq \frac{2\eta_s^2 \mathcal{L}^2(\theta_s) {\gamma^*}^2}{L_w} = \frac{2\alpha_s^2}{L_w} {\gamma^*}^2. \tag{30}
$$

Therefore

$$\mathcal{L}(\theta_{t+1}) \le \mathcal{L}(\theta_0)\exp\Big(-\frac{\gamma^{*2}}{L_w}\sum_{s=0}^{t}\alpha_s\Big) = \mathcal{O}\Big(\exp\big(-\frac{\gamma^{*2}}{L_w}\sqrt{t}\big)\Big). \tag{31}$$

Relabeling $t+1 \to t$ gives the statement. □

◇ *Proof of Theorem D.2-(2):* **Step 1: Margin decomposition.** By the same margin decomposition as in Theorem D.1-(2) (with η_s replacing η and the stability condition verified above):

$$\left\langle \frac{\theta_{t+1}}{\|\theta_{t+1}\|}, y_i x_i \right\rangle \ge \sqrt{\frac{\mu_w}{L_w}}\gamma^* - \frac{\log(n\mathcal{L}(\theta_0))}{\|\theta_{t+1}\|} - \sqrt{\frac{\mu_w}{L_w}}\gamma^*\frac{E_{t+1}}{P_{t+1}}. \tag{32}$$

We bound each error term to obtain the $\mathcal{O}(\log T/\sqrt{T})$ gap.

Step 2: Norm growth from loss decay. From part (1) (already proved above), $\mathcal{L}(\theta_t) \le C\exp(-c\sqrt{t})$ where $c = \gamma^{*2}/L_w$. Since $\mathcal{L}(\theta_t) \ge \frac{1}{n}e^{-R_*\|\theta_t\|}$, we obtain

$$\|\theta_t\| \ge \frac{1}{R_*}\big(c\sqrt{t} - \log(nC)\big) = \Omega(\sqrt{t}). \tag{33}$$

In particular, $\frac{\log n + \log \mathcal{L}(\theta_0)}{\|\theta_{t+1}\|} = O(t^{-1/2})$.

Step 3: Error term decay. *Step-length lower bound.* From (30), $D_w(\theta_s, \theta_{s+1}) \ge \frac{2\alpha_s^2\gamma^{*2}}{L_w}$. Since w is L_w-smooth, $D_w(\theta_s,\theta_{s+1}) \le \frac{L_w}{2}\|\theta_{s+1}-\theta_s\|^2$, giving $\|\theta_{s+1}-\theta_s\| \ge \frac{2\alpha_s\gamma^*}{L_w}$, so $\sum_{s=0}^{t}\|\theta_{s+1}-\theta_s\| = \Omega(\sqrt{t})$.

Step-length upper bound. By the MD optimality condition $\nabla w(\theta_{s+1}) - \nabla w(\theta_s) = -2\eta_s\nabla\mathcal{L}(\theta_s)$ and $(1/\mu_w)$-smoothness of w^*: $\|\theta_{s+1}-\theta_s\| \le \frac{2\eta_s R_*}{\mu_w}\mathcal{L}(\theta_s) = \frac{2R_*}{\mu_w}\alpha_s$. Using $\eta_s {R_*}^2\mathcal{L}(\theta_s)/\mu_w = \alpha_s {R_*}^2/\mu_w \le 1/2$:

$$\frac{\sum_{s=0}^{t}\frac{\mathcal{L}(\theta_s)\eta_s {R_*}^2}{\mu_w}\|\theta_{s+1}-\theta_s\|}{\sum_{s=0}^{t}\|\theta_{s+1}-\theta_s\|} \le \frac{{R_*}^3 L_w}{\gamma^*\mu_w^2}\cdot\frac{\sum_{s=0}^{t}\alpha_s^2}{\sum_{s=0}^{t}\alpha_s} = \mathcal{O}\left(\frac{\log t}{\sqrt{t}}\right).$$

Step 4: Combine. From Steps 2–3, both error terms in (32) are bounded: $1/\|\theta_{t+1}\| = O(t^{-1/2})$ by (33), and $E_{t+1}/P_{t+1} = \mathcal{O}(\log t/\sqrt{t})$. Therefore the margin gap is $\mathcal{O}(\log(T+1)/\sqrt{T+1})$. Solving for the gap to drop below ϵ gives $T_\epsilon = \widetilde{\mathcal{O}}(1/\epsilon^2)$. □

Proof of Corollary D.3

Proof. When $w(\cdot) = \langle\cdot, A\cdot\rangle$ with $\|\cdot\| = \|\cdot\|_A$, we have $L_w = \mu_w = 2$, so $\sqrt{\mu_w/L_w} = 1$. Theorem D.1 then gives $\liminf_{t\to\infty}\min_{i\in[n]}\langle\theta_t/\|\theta_t\|, y_i x_i\rangle \ge \gamma^*$. Since γ^* is the maximum $\|\cdot\|_*$-margin, the reverse inequality $\min_i\langle u, y_i x_i\rangle \le \gamma^*$ holds for any unit vector u, so the limit exists and equals γ^*. Directional convergence $\theta_t/\|\theta_t\| \to u^*$ follows by the same uniqueness argument as in the proof of Corollary 3.7: the max-margin direction is the unique maximizer of the concave, positively homogeneous function $\phi(u) = \min_i\langle u, y_i x_i\rangle$ over $\{u : \|u\| \le 1\}$. Finally, the equivalence to steepest descent: with $w(\cdot) = \langle\cdot, A\cdot\rangle$, the MD update becomes $\nabla w(\theta_{t+1}) = \nabla w(\theta_t) - 2\eta_t\nabla\mathcal{L}(\theta_t)$, i.e., $A\theta_{t+1} = A\theta_t - \eta_t\nabla\mathcal{L}(\theta_t)$, which is steepest descent in the A-metric. □

Exact sharpness for MD

Proposition D.5 (Exact sharpness for MD)**.** *Let $w(\theta) = \frac{1}{2}\theta^\top A\theta$, let $\|\cdot\|$ be any norm, and let Δ° and the one-point dataset $\mathcal{S} = \{(A\Delta^\circ, +1)\}$ be as in Theorem 3.8. For every constant stepsize $\eta > 0$ and $\theta_0 = 0$, the MD iterates satisfy $\theta_t = \rho_t\Delta^\circ$ with $\rho_t \uparrow \infty$, and the margin ratio equals c_w for every $t \ge 1$.*

Proof. Let $z = A\Delta^\circ$, so $\mathcal{L}(\theta) = e^{-\langle\theta, z\rangle}$ and $\nabla\mathcal{L}(\theta) = -e^{-\langle\theta,z\rangle}z$. The MD optimality condition is $A(\theta_{t+1}-\theta_t) = 2\eta e^{-\langle\theta_t, z\rangle}z$, i.e. $\theta_{t+1}-\theta_t = 2\eta e^{-\langle\theta_t,z\rangle}\Delta^\circ$. From $\theta_0 = 0$, induction gives $\theta_t = \rho_t\Delta^\circ$

with $\rho_0 = 0$ and $\rho_{t+1} = \rho_t + 2\eta e^{-\rho_t \langle \Delta^\circ, z\rangle}$. The increments are positive, so ρ_t increases; if ρ_t were bounded, its increments would be bounded below by a positive constant, a contradiction, so $\rho_t \uparrow \infty$. For $t \geq 1$ the direction $\theta_t / \|\theta_t\| = \Delta^\circ / \|\Delta^\circ\|$ is the same as for BPPA, and the computation in the proof of Theorem 3.8 gives the margin ratio $\langle A\Delta^\circ, \Delta^\circ\rangle / (\|\Delta^\circ\| \, \|A\Delta^\circ\|_*) = c_w$. No stability condition is needed because the iterates stay on a ray. □

Remark D.6 (Alignment constant for MD). In the proof of Theorem D.1-(2) the only use of Condition 2 beyond the stability step is the lower bound $D_w(\theta_s, \theta_{s+1}) + D_w(\theta_{s+1}, \theta_s) = \langle \Delta_s^*, \Delta_s \rangle \geq c_s \, \|\Delta_s\| \, \|\Delta_s^*\|_*$ with $c_s \geq \sqrt{\mu_w / L_w}$; replacing this by $c_s \geq c_w$ under Condition 3 gives the margin bound $c_w \gamma^*$ verbatim, as long as the stability condition (which uses strong convexity) is retained.

E Proofs in Section 4

Relation to boosting. KL-BPPA differs from the greedy coordinate steps of boosting, for which the ℓ_1-margin is classical (Rosset et al., 2004; Telgarsky, 2013): all coordinates are updated at once, and the step is implicit.

Reading guide. We first prove the equivalence between KL-BPPA on the doubled variables and hyperbolic-entropy BPPA (Lemma 4.1), then prove the elementary properties of w_β that the analysis uses (Lemma E.1), derive the loss rate, the a priori norm bounds and the vanishing of the steps (Proposition E.2), and finally prove Theorem 4.2. Throughout, $z_i = y_i x_i$, Z has rows $z_i^\top$, $q(\theta) \in \Delta_n$ denotes the loss weights $q_i(\theta) \propto e^{-\langle \theta, z_i \rangle}$, so that

$$-\nabla \mathcal{L}(\theta) = \mathcal{L}(\theta) \, Z^\top q(\theta), \qquad \left\| Z^\top q \right\|_\infty \leq R_\infty \;\text{ and }\; \left\langle u^*, Z^\top q \right\rangle \geq \gamma_1 \quad \text{for every } q \in \Delta_n, \tag{34}$$

where the last inequality holds because $\langle u^*, z_i \rangle \geq \gamma_1$ for every i and q is a probability vector.

Proof of Lemma 4.1. Existence and uniqueness. The objective in (7) is continuous on the closed orthant $\mathbb{R}^{2d}_+$ (with the convention $0 \log 0 = 0$), strictly convex (the two KL terms are jointly strictly convex in (a, b), and $\mathcal{L}(a - b)$ is convex), and coercive (the generalized KL divergence $\mathrm{KL}(a \| a_t) = \sum_i [a_i \log(a_i / a_{t,i}) - a_i + a_{t,i}]$ grows superlinearly in $\|a\|$, while $\mathcal{L} \geq 0$). Hence a unique minimizer exists on the closed orthant. It lies in the open orthant: at a point with $a_i = 0$, the one-sided derivative of $a_i \log a_i$ is $-\infty$ while all other terms have finite derivatives, so moving into the interior decreases the objective. The minimizer is therefore characterized by the stationarity conditions

$$\log \frac{a_{t+1}}{a_t} = -2\eta \, g_{t+1}, \qquad \log \frac{b_{t+1}}{b_t} = 2\eta \, g_{t+1}, \qquad g_{t+1} := \nabla \mathcal{L}(a_{t+1} - b_{t+1}), \tag{35}$$

understood componentwise (note $\nabla_a \mathcal{L}(a - b) = -\nabla_b \mathcal{L}(a - b) = \nabla \mathcal{L}(a - b)$).

Invariance. Adding the two equations in (35) gives $\log(a_{t+1,i} b_{t+1,i}) = \log(a_{t,i} b_{t,i})$, hence $a_{t,i} b_{t,i} = a_{0,i} b_{0,i} = \beta_i^2$ for all t by induction.

Dual variable. Define $\psi_t := \frac{1}{2} \log(a_t / b_t)$. From $a_t b_t = \beta^2$ we get $a_t = \beta e^{\psi_t}$ and $b_t = \beta e^{-\psi_t}$, hence $\theta_t = a_t - b_t = 2\beta \sinh \psi_t$ and $\psi_t = \operatorname{arcsinh}(\theta_t / 2\beta)$. Subtracting the equations in (35) gives $\psi_{t+1} - \psi_t = -2\eta g_{t+1}$.

Identification with w_β. Differentiating $\varphi_\beta(u) := u \operatorname{arcsinh} \frac{u}{2\beta} - \sqrt{u^2 + 4\beta^2}$ gives $\varphi_\beta'(u) = \operatorname{arcsinh} \frac{u}{2\beta} + \frac{u}{\sqrt{u^2 + 4\beta^2}} - \frac{u}{\sqrt{u^2 + 4\beta^2}} = \operatorname{arcsinh} \frac{u}{2\beta}$ and $\varphi_\beta''(u) = (u^2 + 4\beta^2)^{-1/2} > 0$. Hence $\nabla w_\beta(\theta) = \operatorname{arcsinh}(\theta / 2\beta) = \psi$ and $\nabla^2 w_\beta = \operatorname{diag}((\theta_i^2 + 4\beta_i^2)^{-1/2})$. Consequently $\psi_{t+1} - \psi_t = -2\eta \nabla \mathcal{L}(\theta_{t+1})$ is exactly the optimality condition (4) of BPPA instantiated by w_β. Since w_β is strictly convex and supercoercive ($w_\beta(\theta) \geq \sum_i |\theta_i| \log(|\theta_i| / 2\beta_i) - \sqrt{\theta_i^2 + 4\beta_i^2} \to \infty$ faster than linearly), its proximal step has a unique solution, which must therefore equal $a_{t+1} - b_{t+1}$. Finally $\theta_0 = a_0 - b_0 = 0$. □

Lemma E.1 (Properties of w_β). *Let $\beta_{\min} = \min_i \beta_i$ and φ_β as above.*

(a) w_β is $\frac{1}{2\beta_{\min}}$-smooth w.r.t. $\|\cdot\|_2$, is not strongly convex, and its alignment constant is $c_{w_\beta} = 0$ for every norm.

(b) $D_{w_\beta}(\theta,0)=\sum_i\big[\theta_i\operatorname{arcsinh}\frac{\theta_i}{2\beta_i}-\sqrt{\theta_i^2+4\beta_i^2}+2\beta_i\big]\le\|\theta\|_1\log\big(1+\|\theta\|_\infty/\beta_{\min}\big)$.

(c) For all $a,b\in\mathbb{R}$: $D_{\varphi_\beta}(a,b)\ge|b|-|a|\big(1+\log(1+|b|/\beta)\big)-2\beta$; *hence* $D_{w_\beta}(x,y)\ge\|y\|_1-\|x\|_1\big(1+\log(1+\|y\|_\infty/\beta_{\min})\big)-2\sum_i\beta_i$.

(d) If $\psi'=\psi+\delta$ *componentwise,* $\theta=2\beta\sinh\psi$ *and* $\theta'=2\beta\sinh\psi'$, *then for each* i, $|\theta_i'-\theta_i|=\varrho_i\sqrt{\theta_i^2+4\beta_i^2}\,|\delta_i|$ *with* $\varrho_i\in[e^{-|\delta_i|},e^{|\delta_i|}]$; *in particular* $|\theta_i'-\theta_i|\le(\max(|\theta_i|,|\theta_i'|)+2\beta_i)|\delta_i|$ *and* $\|\theta'-\theta\|_1\le e^{\|\delta\|_\infty}\|\delta\|_\infty\sum_i2\beta_i\cosh\psi_i$.

Proof. (a) The Hessian is diagonal with entries $(\theta_i^2+4\beta_i^2)^{-1/2}\in(0,1/(2\beta_i)]$, which tend to 0 as $|\theta_i|\to\infty$; $c_{w_\beta}=0$ is Example C.10(b). (b) Since $\nabla w_\beta(0)=0$ and $w_\beta(0)=-2\sum_i\beta_i$, $D_{w_\beta}(\theta,0)=w_\beta(\theta)-w_\beta(0)$, which is the displayed sum. For the bound use $\operatorname{arcsinh}x\le\log(1+2x)$ for $x\ge0$ (because $\sqrt{1+x^2}\le1+x$) and $2\beta_i-\sqrt{\theta_i^2+4\beta_i^2}\le0$. (c) A direct computation from the definition gives

$$D_{\varphi_\beta}(a,b)=\sqrt{b^2+4\beta^2}-\sqrt{a^2+4\beta^2}+a\Big[\operatorname{arcsinh}\frac{a}{2\beta}-\operatorname{arcsinh}\frac{b}{2\beta}\Big].$$

Now $\sqrt{b^2+4\beta^2}\ge|b|$, $\sqrt{a^2+4\beta^2}\le|a|+2\beta$, $a\operatorname{arcsinh}\frac{a}{2\beta}\ge0$ (same sign), and $-a\operatorname{arcsinh}\frac{b}{2\beta}\ge-|a|\operatorname{arcsinh}\frac{|b|}{2\beta}\ge-|a|\log(1+|b|/\beta)$. Summing over coordinates and using $\|y_i\|\le\|y\|_\infty$ gives the vector statement. (d) By the mean value theorem, $\sinh\psi_i'-\sinh\psi_i=\cosh(\xi_i)\delta_i$ for some ξ_i between ψ_i and ψ_i', and $\cosh(\psi+\delta)=\cosh\psi\cosh\delta+\sinh\psi\sinh\delta\in[e^{-|\delta|}\cosh\psi,e^{|\delta|}\cosh\psi]$ because $|\sinh\psi|\le\cosh\psi$ and $\cosh\delta\pm|\sinh\delta|=e^{\pm|\delta|}$. Finally $2\beta_i\cosh\psi_i=\sqrt{\theta_i^2+4\beta_i^2}\le|\theta_i|+2\beta_i$, and $\cosh(\xi_i)\le\cosh(\max(|\psi_i|,|\psi_i'|))$ gives the version with $\max(|\theta_i|,|\theta_i'|)$. □

Proposition E.2 (Loss rate, norm bounds and vanishing steps for KL-BPPA)**.** *For constant* η *and* $2\eta t\ge1$,

(i) $\mathcal{L}(\theta_t)\le\dfrac{1+\frac{\log(2\eta t)}{\gamma_1}\log\Big(1+\frac{\log(2\eta t)}{\gamma_1\beta_{\min}}\Big)}{2\eta t}=\mathcal{O}\Big(\dfrac{\log t\,\log\log t}{t}\Big)$;

(ii) $\dfrac{1}{R_\infty}\log\dfrac{1}{n\mathcal{L}(\theta_t)}\le\|\theta_t\|_1\le\max\Big\{2A_t,\ 4s_t\log\Big(1+\dfrac{4s_t}{\beta_{\min}}\Big)\Big\}=\mathcal{O}(\log t\,\log\log t)$, *where* $s_t=\log t/\gamma_1$ *and* $A_t=s_t\log(1+s_t/\beta_{\min})+2\eta+2\sum_i\beta_i+s_t$;

(iii) $\|\Delta_t\|_1\le2\eta R_\infty\mathcal{L}(\theta_{t+1})\big(\|\theta_t\|_1+\|\theta_{t+1}\|_1+2\sum_i\beta_i\big)=\mathcal{O}\big(\log^2t(\log\log t)^2/t\big)$; *hence* $\sum_t\|\Delta_t\|_1^2<\infty$.

Proof. (i) By the comparison bound of Lemma C.4 (which uses only the three-point inequality and monotonicity, and holds for any convex differentiable DGF with a well-posed proximal step; its statement with L_w is replaced by $D_w(\theta,\theta_0)$), $\mathcal{L}(\theta_t)\le\mathcal{L}(su^*)+D_{w_\beta}(su^*,0)/(2\eta t)$ for every $s\ge0$. Here $\mathcal{L}(su^*)\le e^{-s\gamma_1}$ and, by Lemma E.1(b) with $\|u^*\|_1=1\ge\|u^*\|_\infty$, $D_{w_\beta}(su^*,0)\le s\log(1+s/\beta_{\min})$. Take $s=\log(2\eta t)/\gamma_1$.

(ii) The lower bound is $\mathcal{L}(\theta_t)\ge\frac{1}{n}e^{-\min_i\langle\theta_t,z_i\rangle}\ge\frac{1}{n}e^{-R_\infty\|\theta_t\|_1}$. For the upper bound, sum the three-point inequality (12) with $\theta=s_tu^*$ over the first t steps:

$$\begin{aligned}D_{w_\beta}(s_tu^*,\theta_t)&\le D_{w_\beta}(s_tu^*,0)+2\eta\sum_{k<t}\big[\mathcal{L}(s_tu^*)-\mathcal{L}(\theta_{k+1})\big]\\&\le s_t\log(1+s_t/\beta_{\min})+2\eta t\,e^{-s_t\gamma_1}=s_t\log(1+s_t/\beta_{\min})+2\eta.\end{aligned}$$

By Lemma E.1(c) with $x=s_tu^*$, $y=\theta_t$, and $\|\theta_t\|_\infty\le\|\theta_t\|_1$, $D_{w_\beta}(s_tu^*,\theta_t)\ge\|\theta_t\|_1-s_t\big(1+\log(1+\|\theta_t\|_1/\beta_{\min})\big)-2\sum_i\beta_i$. Combining, $x:=\|\theta_t\|_1$ satisfies $x\le A_t+s_t\log(1+x/\beta_{\min})$. We claim that $x\le A+B\log(1+x/\beta)$ with $A,B\ge0$ implies $x\le\max\{2A,4B\log(1+4B/\beta)\}$: if $x>2A$ then $x\le2B\log(1+x/\beta)$; with $y=x/\beta$ and $K=2B/\beta$ this becomes $y\le K\log(1+y)$, which forces $y=0$ if $K<1$ and, if $K\ge1$, $y\le2K\log(1+2K)$ because $y\mapsto y/\log(1+y)$

is increasing and $2K\log(1+2K)/\log(1+2K\log(1+2K)) \geq K$ (as $1+2K\log(1+2K) \leq (1+2K)^2$). This gives the bound with $A = A_t$, $B = s_t$.

(iii) By Lemma E.1(d) with $\delta = \psi_{t+1} - \psi_t = -2\eta\nabla\mathcal{L}(\theta_{t+1})$ and $\|\delta\|_\infty \leq 2\eta R_\infty \mathcal{L}(\theta_{t+1})$ (by (34)), summing $|\Delta_{t,i}| \leq (\max(|\theta_{t,i}|, |\theta_{t+1,i}|) + 2\beta_i)|\delta_i|$ over i gives the bound; (i) and (ii) give the rate, whose square is summable. □

PROOF OF THEOREM 4.2

Proof. Throughout, $s_t := \psi_{t+1} - \psi_t = -2\eta\nabla\mathcal{L}(\theta_{t+1}) = \tau_t v_{t+1}$ with $v_{t+1} := Z^\top q(\theta_{t+1})$ and $\tau_t = 2\eta\mathcal{L}(\theta_{t+1})$, so that $\|s_t\|_\infty \leq \tau_t R_\infty$ by (34); and $P_t = \sum_i (a_{t,i} + b_{t,i}) = \sum_i 2\beta_i \cosh\psi_{t,i}$, so that $\|\theta_t\|_1 \leq P_t \leq \|\theta_t\|_1 + 2\sum_i \beta_i$ (because $|\sinh| \leq \cosh \leq |\sinh| + 1$) and $P_t \geq 2\sum_i \beta_i$.

Step 1 (the multiplicative-weights potential). Since $a_{t+1,i} = a_{t,i}e^{s_{t,i}}$ and $b_{t+1,i} = b_{t,i}e^{-s_{t,i}}$ by (35),

$$\log P_{t+1} - \log P_t = \log \frac{\sum_i \left(a_{t,i}e^{s_{t,i}} + b_{t,i}e^{-s_{t,i}}\right)}{P_t} = \log \mathbb{E}_{\pi_t}\left[e^{X_t}\right],$$

where π_t is the probability distribution on $2d$ points that puts mass $a_{t,i}/P_t$ on the value $X_t = s_{t,i}$ and mass $b_{t,i}/P_t$ on the value $X_t = -s_{t,i}$. The random variable X_t takes values in $[-\|s_t\|_\infty, \|s_t\|_\infty]$, so Hoeffding's lemma ($\log \mathbb{E}e^X \leq \mathbb{E}X + (\text{range})^2/8$) gives

$$\log P_{t+1} - \log P_t \leq \mathbb{E}_{\pi_t} X_t + \frac{\|s_t\|_\infty^2}{2}, \qquad \mathbb{E}_{\pi_t} X_t = \frac{\sum_i (a_{t,i} - b_{t,i})s_{t,i}}{P_t} = \frac{\langle\theta_t, s_t\rangle}{P_t} = \frac{\tau_t \langle\theta_t, v_{t+1}\rangle}{P_t}.$$

Summing over $t < T$ and using $P_0 = 2\sum_i \beta_i$ and $\|s_t\|_\infty \leq \tau_t R_\infty$,

$$\log P_T \leq \log\left(2\sum_i \beta_i\right) + \sum_{t<T} \frac{\tau_t \langle\theta_t, v_{t+1}\rangle}{P_t} + \frac{{R_\infty}^2}{2}\sum_{t<T}\tau_t^2. \tag{36}$$

Step 2 (the best expert). For each coordinate i, $P_T \geq a_{T,i} + b_{T,i} \geq \max(a_{T,i}, b_{T,i}) = \beta_i e^{|\psi_{T,i}|}$, hence

$$\log P_T \geq \|\psi_T\|_\infty + \log\beta_{\min}. \tag{37}$$

Moreover, since $\|u^*\|_1 = 1$, Hölder's inequality and (34) give

$$\|\psi_T\|_\infty \;\geq\; \langle\psi_T, u^*\rangle = \sum_{t<T} \tau_t \langle u^*, v_{t+1}\rangle \;\geq\; \gamma_1 \sum_{t<T} \tau_t = \gamma_1 A_T, \tag{38}$$

where we used $\psi_0 = 0$ and $\psi_T = \sum_{t<T} s_t$.

Step 3 (the reward is at most the current margin). By (34), $\langle\theta_t, v_{t+1}\rangle = \sum_j q_{t+1,j}\langle\theta_t, z_j\rangle = \mathbb{E}_{q_{t+1}}\langle\theta_t, z\rangle \leq \mathbb{E}_{q_{t+1}}\langle\theta_{t+1}, z\rangle + R_\infty\|\Delta_t\|_1$, since $|\langle\Delta_t, z_j\rangle| \leq \|\Delta_t\|_1\|z_j\|_\infty$. For $q = q(\theta)$ and $x_j := \langle\theta, z_j\rangle$, the Gibbs variational identity $H(q) = \mathbb{E}_q x + \log\sum_j e^{-x_j}$ (which follows from $\log q_j = -x_j - \log\sum_k e^{-x_k}$) gives

$$\mathbb{E}_{q(\theta)}\langle\theta, z\rangle = H(q(\theta)) - \log\sum_j e^{-x_j} = H(q(\theta)) - \log n - \log\mathcal{L}(\theta) \leq \log n + m(\theta),$$

using $H(q) \leq \log n$ and $\mathcal{L}(\theta) \geq \frac{1}{n}e^{-m(\theta)}$, where $m(\theta) = \min_j\langle\theta, z_j\rangle$. Hence

$$\langle\theta_t, v_{t+1}\rangle \leq m(\theta_{t+1}) + \log n + R_\infty\|\Delta_t\|_1. \tag{39}$$

Write $m(\theta_{t+1}) = \mu_1(\theta_{t+1})\|\theta_{t+1}\|_1$ and note $P_t - 2\sum_i\beta_i - \|\Delta_t\|_1 \leq \|\theta_t\|_1 - \|\Delta_t\|_1 \leq \|\theta_{t+1}\|_1 \leq P_t + \|\Delta_t\|_1$ and $|\mu_1(\theta)| \leq R_\infty$. If $\mu_1(\theta_{t+1}) \geq 0$ then $m(\theta_{t+1}) \leq \mu_1(\theta_{t+1})P_t + R_\infty\|\Delta_t\|_1$; if $\mu_1(\theta_{t+1}) < 0$ then $m(\theta_{t+1}) \leq \mu_1(\theta_{t+1})(P_t - 2\sum\beta_i - \|\Delta_t\|_1) \leq \mu_1(\theta_{t+1})P_t + R_\infty(2\sum_i\beta_i + \|\Delta_t\|_1)$. In both cases

$$\frac{m(\theta_{t+1})}{P_t} \leq \mu_1(\theta_{t+1}) + \frac{R_\infty\left(\|\Delta_t\|_1 + 2\sum_i\beta_i\right)}{P_t}. \tag{40}$$

Step 4 (combining). Insert (39) and (40) into (36), and use (37)–(38) on the left:

$$\sum_{t<T}\tau_t\big(\gamma_1 - \mu_1(\theta_{t+1})\big) \;\leq\; \log\frac{2\sum_i\beta_i}{\beta_{\min}} + \Big(\log n + 2R_\infty\sum_i\beta_i\Big)\sum_{t<T}\frac{\tau_t}{P_t}$$

$$+ 2R_\infty \sum_{t<T} \frac{\tau_t \|\Delta_t\|_1}{P_t} + \frac{{R_\infty}^2}{2} \sum_{t<T} \tau_t^2. \tag{41}$$

We show that the three series on the right converge, uniformly in T. (i) $\sum_t \tau_t^2 = 4\eta^2 \sum_t \mathcal{L}(\theta_{t+1})^2 < \infty$ by Proposition E.2(i). (ii) By Lemma E.1(d) with $\delta = s_t$, $\|\Delta_t\|_1 \le e^{\|s_t\|_\infty} \|s_t\|_\infty P_t \le e^{\tau_t R_\infty} \tau_t R_\infty P_t \le e^{2\eta R_\infty} R_\infty \tau_t P_t$, since $\tau_t \le 2\eta \mathcal{L}(\theta_0) = 2\eta$. Hence $\sum_t \tau_t \|\Delta_t\|_1 / P_t \le e^{2\eta R_\infty} R_\infty \sum_t \tau_t^2 < \infty$. (iii) By (37)–(38), $P_t \ge \beta_{\min} e^{\|\psi_t\|_\infty} \ge \beta_{\min} e^{\gamma_1 A_t}$. Since $A_{t+1} = A_t + \tau_t$ and $\tau_t \le 2\eta$,

$$\sum_t \frac{\tau_t}{P_t} \le \frac{1}{\beta_{\min}} \sum_t \tau_t e^{-\gamma_1 A_t} \le \frac{e^{2\eta\gamma_1}}{\beta_{\min}} \sum_t \tau_t e^{-\gamma_1 A_{t+1}} \le \frac{e^{2\eta\gamma_1}}{\beta_{\min}} \sum_t \int_{A_t}^{A_{t+1}} e^{-\gamma_1 r} dr \le \frac{e^{2\eta\gamma_1}}{\gamma_1 \beta_{\min}},$$

where we used $\tau_t e^{-\gamma_1 A_{t+1}} \le \int_{A_t}^{A_{t+1}} e^{-\gamma_1 r} dr$ (the integrand is at least $e^{-\gamma_1 A_{t+1}}$ on the interval). Collecting the three bounds gives statement (a) with the explicit constant

$$C_\infty = \log \frac{2\sum_i \beta_i}{\beta_{\min}} + \frac{(\log n + 2R_\infty \sum_i \beta_i) e^{2\eta\gamma_1}}{\gamma_1 \beta_{\min}} + \Big(2{R_\infty}^2 e^{2\eta R_\infty} + \frac{{R_\infty}^2}{2}\Big) \sum_{t\ge 0} \tau_t^2;$$

the summands on the left are non-negative because $\mu_1 \le \gamma_1$ always.

For (b), the same chain without the last step gives $\log P_T \le \log(2\sum \beta_i) + \sum_{t<T} \tau_t \mu_1(\theta_{t+1}) +$ (the three error sums) $\le \log(2\sum \beta_i) + \gamma_1 A_T + (\dots)$, and with (37), $\|\psi_T\|_\infty - \langle \psi_T, u^* \rangle \le \log P_T - \log \beta_{\min} - \gamma_1 A_T \le C_\infty$. The coordinatewise statement follows from $\|\psi\|_\infty - \langle \psi, u^* \rangle = \sum_i |u_i^*| \big(\|\psi\|_\infty - \operatorname{sign}(u_i^*) \psi_i \big)$, a sum of non-negative terms (recall $\|u^*\|_1 = 1$), each of which is therefore at most C_∞.

Step 5 (from an integrated to a pointwise statement). Let $f_t := \gamma_1 - \mu_1(\theta_{t+1}) \ge 0$. Since $m(\cdot)$ is R_∞-Lipschitz w.r.t. $\|\cdot\|_1$ and $|m(\theta)| \le R_\infty \|\theta\|_1$,

$$\begin{aligned} |\mu_1(\theta_{t+1}) - \mu_1(\theta_t)| &= \Big| \frac{m(\theta_{t+1}) - m(\theta_t)}{\|\theta_{t+1}\|_1} + m(\theta_t) \Big(\frac{1}{\|\theta_{t+1}\|_1} - \frac{1}{\|\theta_t\|_1} \Big) \Big| \\ &\le \frac{R_\infty \|\Delta_t\|_1}{\|\theta_{t+1}\|_1} + \frac{R_\infty \|\theta_t\|_1 \|\Delta_t\|_1}{\|\theta_t\|_1 \|\theta_{t+1}\|_1} = \frac{2R_\infty \|\Delta_t\|_1}{\|\theta_{t+1}\|_1}. \end{aligned}$$

For t large enough, $\|\theta_{t+1}\|_1 \ge P_t - 2\sum \beta_i - \|\Delta_t\|_1 \ge P_t/2$, because $P_t \ge \|\theta_t\|_1 \to \infty$ and $\|\Delta_t\|_1 / P_t \le e^{2\eta R_\infty} R_\infty \tau_t \to 0$. Hence, by (ii) above,

$$|f_{t+1} - f_t| \le C_L \, \tau_{t+1}, \qquad C_L := 4{R_\infty}^2 e^{2\eta R_\infty}.$$

Also $\tau_t \to 0$ and $A_T \to \infty$: indeed $R_\infty A_T \ge \sum_{t<T} \|s_t\|_\infty \ge \|\psi_T\|_\infty \ge \log(P_T/2\sum \beta_i) \to \infty$, using (37) in the form $P_T \le 2\sum_i \beta_i e^{\|\psi_T\|_\infty}$ and $P_T \ge \|\theta_T\|_1 \to \infty$. Suppose $f_t \not\to 0$. Then there are $\varepsilon > 0$ and infinitely many indices t_k with $f_{t_k} \ge \varepsilon$; thin them so that the "clock intervals" $I_k := \{t \ge t_k : A_{t+1} - A_{t_k+1} \le \varepsilon/(2C_L)\}$ are pairwise disjoint (possible since $A_T \to \infty$). For $t \in I_k$, $f_t \ge f_{t_k} - C_L \sum_{s=t_k+1}^{t} \tau_s = f_{t_k} - C_L(A_{t+1} - A_{t_k+1}) \ge \varepsilon/2$. Moreover $\sum_{t\in I_k} \tau_t \ge \varepsilon/(2C_L) - \sup_{t\ge t_k} \tau_t \ge \varepsilon/(4C_L)$ for k large (the interval is maximal, so adding one more term would exceed $\varepsilon/(2C_L)$). Hence $\sum_{t\in I_k} \tau_t f_t \ge \varepsilon^2/(8C_L)$ for infinitely many disjoint I_k, contradicting (a). Therefore $f_t \to 0$, which is (c).

Step 6 (direction). Let $\bar{u}$ be a limit point of $\theta_t / \|\theta_t\|_1$. Then $\|\bar{u}\|_1 = 1$ and, since $u \mapsto \min_i \langle u, z_i \rangle$ is continuous, $\min_i \langle \bar{u}, z_i \rangle = \lim \mu_1(\theta_t) = \gamma_1$ by (c), so $\bar{u}$ is a maximum ℓ_1-margin classifier. If the maximizer is unique, every limit point of the bounded sequence $\theta_t / \|\theta_t\|_1$ equals u^*, hence the sequence converges to u^*. □

Remark E.3 (Adaptive stepsizes). For $\eta_t = \alpha_t / \mathcal{L}(\theta_t)$ one has $\tau_t = 2\alpha_t r_t$ with $r_t = \mathcal{L}(\theta_{t+1})/\mathcal{L}(\theta_t)$. The proof above uses two facts that follow from the loss rate of Proposition E.2 for constant η: $\sum_t \tau_t^2 < \infty$ and $A_T \to \infty$. Under the adaptive schedule both require a lower bound on r_t, which the entropic geometry does not provide, so we leave the adaptive case of Theorem 4.2 open.

F PROOFS IN SECTION 5

Reading guide. We first prove Lemma 5.1, which identifies the prediction-space KL with D_W and shows why the geometry is degenerate. We then prove the logit reduction (Lemma 5.2), solve the scalar proximal map (Lemma F.1), and prove Theorem 5.3 part by part. The section ends with the squared prediction distance (Proposition F.2). Throughout, $z_i = y_i x_i$, $s_i(\theta) = \langle \theta, z_i \rangle$, $u_i = e^{-s_i}$, and Φ denotes the proximal objective written in logit space. Condition 1 implies $z_i \neq 0$ for every i.

PROOF OF LEMMA 5.1

Proof. **Step 1 (KL is a Bregman divergence).** With $p = \sigma(s)$ one has $\log p = s - A(s)$ and $\log(1-p) = -A(s)$. Expanding, with $p' = \sigma(s')$,

$$\mathrm{KL}(\mathrm{Bern}(p') \| \mathrm{Bern}(p)) = p' \log \frac{p'}{p} + (1-p') \log \frac{1-p'}{1-p} = A(s) - A(s') - p'(s-s') = D_A(s, s').$$

For the linear map $\theta \mapsto \langle \theta, z_i \rangle$, $\nabla(A \circ \langle \cdot, z_i \rangle)(\theta') = \sigma(s_i(\theta'))z_i$, so $D_{A \circ \langle \cdot, z_i \rangle}(\theta, \theta') = D_A(s_i(\theta), s_i(\theta'))$. Averaging over i gives (8); the identity $D_A(-s, -s') = D_A(s, s')$, which follows from $A(-s) = A(s) - s$ and $\sigma(-s') = 1 - \sigma(s')$, lets us fold the labels into z_i.

Step 2 (part (i)). W is convex because $A'' = \sigma' > 0$ and W is an average of convex functions composed with linear maps. Let $v, v' \in \mathrm{null}(Z)$. Then $s_i(\theta + v) = s_i(\theta)$ for every i, so $W(\theta + v) = W(\theta)$ and $\nabla W(\theta + v) = \frac{1}{n} \sum_i \sigma(s_i(\theta)) z_i = \nabla W(\theta)$. Moreover $\nabla W(\theta') \in \mathrm{row}(Z)$, hence $\langle \nabla W(\theta'), v - v' \rangle = 0$. Therefore

$$D_W(\theta + v, \theta' + v') = W(\theta) - W(\theta') - \langle \nabla W(\theta'), \theta - \theta' \rangle - \langle \nabla W(\theta'), v - v' \rangle = D_W(\theta, \theta').$$

The loss also depends on θ only through $Z\theta$, so the proximal objective is $F(\theta) = \Phi(Z\theta)$ with $\Phi(s) := \frac{1}{n} \sum_i \left[\ell(s_i) + \frac{1}{2\eta} D_A(s_i, s_{t,i}) \right]$. Each summand of Φ is strictly convex and coercive: as $s_i \to +\infty$, $D_A(s_i, s_{t,i})$ grows linearly with slope $1 - \sigma(s_{t,i}) > 0$, and as $s_i \to -\infty$, $\ell(s_i) \to \infty$. Hence Φ has a unique minimizer $\bar{s}$ over the subspace $\mathrm{range}(Z)$, and the minimizers of F are exactly $\{\theta : Z\theta = \bar{s}\}$, a nonempty translate of $\mathrm{null}(Z)$. When $\mathrm{rank} Z < d$ this set is not a single point.

Step 3 (reduction to the Euclidean norm). Let $a \|x\|_2 \leq \|x\| \leq b \|x\|_2$. Then $\frac{1}{b} \|g\|_2 \leq \|g\|_* \leq \frac{1}{a} \|g\|_2$, and for every pair $x \neq y$,

$$c_{\|\cdot\|}(x, y) = \frac{\langle \Delta^*, \Delta \rangle}{\|\Delta\| \, \|\Delta^*\|_*} \leq \frac{\langle \Delta^*, \Delta \rangle}{a \|\Delta\|_2 \cdot \frac{1}{b} \|\Delta^*\|_2} = \frac{b}{a} \, c_{\|\cdot\|_2}(x, y).$$

Hence it suffices to prove $c_W = 0$ for $\|\cdot\| = \|\cdot\|_2$, whose dual norm is itself.

Step 4 (part (ii) when $\mathrm{rank} Z < d$). Take $v \in \mathrm{null}(Z) \setminus \{0\}$, h with $Zh \neq 0$, $\varepsilon > 0$, $y = \theta$ and $x = \theta + v + \varepsilon h$. By Step 2, $\Delta^* = \nabla W(x) - \nabla W(y) = \nabla W(\theta + \varepsilon h) - \nabla W(\theta)$, which lies in $\mathrm{row}(Z)$ and is nonzero because $\langle \Delta^*, h \rangle = \varepsilon \int_0^1 h^\top \nabla^2 W(\theta + r\varepsilon h) h \, dr > 0$ (the integrand equals $\frac{1}{n} \sum_i \sigma'(\cdot) \langle z_i, h \rangle^2 > 0$). Since $\langle \Delta^*, v \rangle = 0$ and by the Cauchy–Schwarz inequality,

$$c(x, y) = \frac{\langle \Delta^*, v + \varepsilon h \rangle}{\|v + \varepsilon h\|_2 \, \|\Delta^*\|_2} = \frac{\varepsilon \langle \Delta^*, h \rangle}{\|v + \varepsilon h\|_2 \, \|\Delta^*\|_2} \leq \frac{\varepsilon \|h\|_2}{\|v + \varepsilon h\|_2} \longrightarrow 0 \qquad (\varepsilon \to 0).$$

Geometrically, Δ points mainly along the null-space direction v, while Δ^* is confined to $\mathrm{row}(Z)$.

Step 5 (part (ii) when $\mathrm{rank} Z = d \geq 2$ and no two z_i are parallel). Now $H(y) := \nabla^2 W(y) = \frac{1}{n} Z^\top \mathrm{diag}(\sigma'(s(y))) Z$ is positive definite for every y.

(a) Localization. Fix y and a unit vector e, and let $\Delta = re$ with $r \to 0$. By Taylor's theorem, $\Delta^* = rH(y)e + o(r)$, so $\langle \Delta^*, \Delta \rangle = r^2 \langle H(y)e, e \rangle + o(r^2)$, $\|\Delta\|_2 = r$ and $\|\Delta^*\|_2 = r \|H(y)e\|_2 + o(r)$. Hence $c(y + re, y) \to \langle H(y)e, e \rangle / \|H(y)e\|_2$. Taking the infimum over e and using Lemma B.5 and its equality case,

$$c_W \leq \inf_y \frac{2\sqrt{\kappa_y}}{1 + \kappa_y}, \qquad \kappa_y := \frac{\lambda_{\max}(H(y))}{\lambda_{\min}(H(y))}.$$

It remains to find points y with $\kappa_y \to \infty$.

(b) Choice of direction. The set of unit vectors $\widehat{\theta}$ with $\left\langle \widehat{\theta}, z_i \right\rangle = 0$ for some i, or with $|\left\langle \widehat{\theta}, z_i \right\rangle| = |\left\langle \widehat{\theta}, z_j \right\rangle|$ for some $i \neq j$, is contained in a finite union of hyperplanes, because $z_i \neq 0$ and $z_i \mp z_j \neq 0$ when z_i, z_j are not parallel. Pick $\widehat{\theta}$ outside it and let $a_i := |\left\langle \widehat{\theta}, z_i \right\rangle|$. The a_i are positive and pairwise distinct; order them as $a_{(1)} < a_{(2)} < \cdots < a_{(n)}$, with corresponding samples $z_{(k)}$. Since $n \geq d \geq 2$, $a_{(d)} > a_{(1)}$.

(c) Eigenvalue bounds. Let $y = R\widehat{\theta}$ with $R > 0$ and $R_2 := \max_i \|z_i\|_2$. From $\sigma'(s) = e^{-|s|}/(1 + e^{-|s|})^2$ we get $\frac{1}{4}e^{-|s|} \leq \sigma'(s) \leq e^{-|s|}$, so the weights satisfy $w_i := \sigma'(s_i(y)) \in [\frac{1}{4}e^{-Ra_i}, e^{-Ra_i}]$. With $u = z_{(1)}/\left\|z_{(1)}\right\|_2$, keeping only one non-negative term,

$$\lambda_{\max}(H(y)) \;\geq\; u^\top H(y) u \;\geq\; \frac{1}{n} w_{(1)} \left\|z_{(1)}\right\|_2^2 \;\geq\; \frac{\left\|z_{(1)}\right\|_2^2}{4n} e^{-Ra_{(1)}}.$$

Choose a unit vector u' orthogonal to $z_{(1)}, \ldots, z_{(d-1)}$, which exists because these are $d-1$ vectors in $\mathbb{R}^d$. By the Courant–Fischer theorem,

$$\lambda_{\min}(H(y)) \;\leq\; u'^\top H(y) u' = \frac{1}{n} \sum_{k \geq d} w_{(k)} \left\langle u', z_{(k)} \right\rangle^2 \;\leq\; R_2^2 \, e^{-Ra_{(d)}}.$$

Therefore $\kappa_y \geq \frac{\left\|z_{(1)}\right\|_2^2}{4nR_2^2} e^{R(a_{(d)} - a_{(1)})} \to \infty$ as $R \to \infty$, and (a) gives $c_W = 0$. □

PROOF OF LEMMA 5.2

Proof. By Step 2 above, the proximal objective is $F(\theta) = \Phi(Z\theta)$ with Φ separable, strictly convex and coercive. Since $\mathrm{rank} Z = n$, Z maps $\mathbb{R}^d$ onto $\mathbb{R}^n$, so the unconstrained minimizer $s_{t+1} = \mathrm{argmin}_{s \in \mathbb{R}^n} \Phi(s)$ is attained and the minimizers of F are exactly $\{\theta : Z\theta = s_{t+1}\}$. Because Φ is separable, $s_{t+1,i}$ minimizes the scalar function $s \mapsto \ell(s) + \frac{1}{2\eta} D_A(s, s_{t,i})$, which is the scalar map T_η of Lemma F.1. The gradients $\nabla F(\theta) = Z^\top \nabla \Phi(Z\theta)$ lie in $\mathrm{row}(Z)$, so gradient descent on F started at θ_t stays in the affine space $\theta_t + \mathrm{row}(Z)$. On this space Z is injective, so F restricted to it is strictly convex with a unique minimizer. Writing $\theta = \theta_t + Z^\top c$, the constraint $Z\theta = s_{t+1}$ becomes $Kc = s_{t+1} - s_t$, so the minimizer is $\theta_t + Z^\top K^{-1}(s_{t+1} - s_t) = \theta_t + Z^\dagger(s_{t+1} - s_t)$. On the sublevel set of F in $\theta_t + \mathrm{row}(Z)$, which is compact by coercivity of Φ and injectivity of Z, the function F is smooth and strongly convex, so gradient descent with a sufficiently small stepsize converges to this minimizer. It is also the minimizer closest to θ_t in $\|\cdot\|_2$. Finally, since $Z^\dagger Z$ is the orthogonal projection onto $\mathrm{row}(Z)$ and the increments lie in $\mathrm{row}(Z)$, the null-space component of θ_t equals that of θ_0, which gives $\theta_t = Z^\dagger s_t + (I - Z^\dagger Z)\theta_0$. □

Lemma F.1 (Scalar proximal map)**.** *Let $\eta > 0$, $s \in \mathbb{R}$, $u = e^{-s}$, and let $s' = T_\eta(s)$, $u' = e^{-s'}$. For the exponential loss, s' is the unique solution of $\sigma(s') - \sigma(s) = 2\eta e^{-s'}$, $s' > s$, and*

$$u = u'\big[1 + 2\eta(1+u)(1+u')\big], \qquad \frac{u}{1 + 2\eta(1+u)^2} \leq u' \leq \frac{u}{1+2\eta},$$
$$\log(1+2\eta) < s' - s \leq \log\big(1 + 2\eta(1+u)^2\big).$$

Proof. The first-order condition of $s' \mapsto \ell(s') + \frac{1}{2\eta} D_A(s', s)$ is $\ell'(s') + \frac{1}{2\eta}(\sigma(s') - \sigma(s)) = 0$, because $\partial_{s'} D_A(s', s) = \sigma(s') - \sigma(s)$. With $\ell'(s') = -e^{-s'}$, the function $h(s') = \sigma(s') - \sigma(s) - 2\eta e^{-s'}$ is strictly increasing with $h(s) < 0 < h(+\infty) = 1 - \sigma(s)$, so there is a unique root, and it exceeds s. Substituting $\sigma(s) = 1/(1+u)$,

$$\frac{1}{1+u'} - \frac{1}{1+u} = \frac{u - u'}{(1+u)(1+u')} = 2\eta u',$$

which is the identity. Since $0 < u' < u$, the factor $(1+u)(1+u')$ lies in $(1, (1+u)^2)$, which gives the bounds; the lower bound on $s' - s = \log(u/u')$ is strict because $(1+u)(1+u') > 1$. □

PROOF OF THEOREM 5.3

Proof. Let $u_{t,i} = e^{-s_{t,i}}$, $u_{\max} \coloneqq \max_i u_{0,i}$, and $\varkappa \coloneqq (1 + u_{\max})^2$. By Lemma 5.2 every logit follows $s_{t+1,i} = T_\eta(s_{t,i})$, and by Lemma F.1 the logits increase, so $u_{t,i} \le u_{0,i}$.

(1) Geometric decay. Since $\mathcal{L}(\theta_t) = \frac{1}{n}\sum_i u_{t,i}$ and Lemma F.1 gives $u_{t+1,i} \le u_{t,i}/(1+2\eta)$ for every i, averaging over i gives (1).

(2) Equal logit growth. Fix i and drop it from the notation. Define the excess increment

$$e_k \coloneqq s_{k+1} - s_k - \log(1+2\eta) = \log \frac{1 + 2\eta(1+u_k)(1+u_{k+1})}{1+2\eta}.$$

Since $(1+u_k)(1+u_{k+1}) > 1$, $e_k > 0$. Writing the ratio as $1 + \frac{2\eta}{1+2\eta}[(1+u_k)(1+u_{k+1}) - 1]$ and using $\log(1+x) \le x$,

$$e_k \le \frac{2\eta}{1+2\eta}\big[u_k + u_{k+1}(1+u_k)\big] \le \frac{2\eta}{1+2\eta}(2+u_0)\,u_k \le \frac{2\eta}{1+2\eta}(1+\varkappa)\,u_k,$$

because $u_{k+1} \le u_k \le u_0$. By Lemma F.1, $u_k \le u_0(1+2\eta)^{-k}$. Hence $v \coloneqq \sum_{k\ge 0} e_k$ converges, and with $\varepsilon_t \coloneqq \sum_{k\ge t} e_k$ we have the exact identity

$$s_t = s_0 + t\log(1+2\eta) + \sum_{k<t} e_k = s_0 + t\log(1+2\eta) + v - \varepsilon_t.$$

The quantity ε_t is the remaining excess growth after time t: it is positive, non-increasing, $\varepsilon_0 = v$, and summing the geometric series with $\sum_{k\ge t}(1+2\eta)^{-k} = (1+2\eta)^{1-t}/(2\eta)$,

$$0 < \varepsilon_t \le (1+\varkappa)\,u_0\,(1+2\eta)^{-t}, \qquad 0 < v \le (1+\varkappa)\,u_0.$$

(3) Direction. Let $g \coloneqq Z^\dagger \mathbf{1}$; it is nonzero because $Zg = ZZ^\dagger\mathbf{1} = \mathbf{1}$. Telescoping the selection rule and inserting (2),

$$\theta_t = \theta_0 + Z^\dagger(s_t - s_0) = t\log(1+2\eta)\,g + r_t, \qquad r_t \coloneqq \theta_0 + Z^\dagger(v - \varepsilon_t).$$

Since $0 \le v - \varepsilon_t \le v$ componentwise and $\big\|Z^\dagger\big\|_{\mathrm{op}} = \lambda_{\min}(K)^{-1/2}$,

$$\|r_t\|_2 \le B \coloneqq \|\theta_0\|_2 + \frac{\|v\|_2}{\sqrt{\lambda_{\min}(K)}} \le \|\theta_0\|_2 + \frac{(1+\varkappa)\,\|u_0\|_2}{\sqrt{\lambda_{\min}(K)}},$$

uniformly in t. For nonzero vectors a, b,

$$\Big\|\frac{a}{\|a\|} - \frac{b}{\|b\|}\Big\| = \Big\|\frac{a-b}{\|b\|} + a\Big(\frac{1}{\|a\|} - \frac{1}{\|b\|}\Big)\Big\| \le \frac{\|a-b\|}{\|b\|} + \frac{|\,\|b\| - \|a\|\,|}{\|b\|} \le \frac{2\,\|a-b\|}{\|b\|}.$$

Apply this with $a = \theta_t$ and $b = t\log(1+2\eta)\,g$, which is a positive multiple of g for $t \ge 1$; then $a - b = r_t$, and $\theta_t \ne 0$ as soon as $t\log(1+2\eta)\,\|g\|_2 > B$. Hence

$$\Big\|\frac{\theta_t}{\|\theta_t\|_2} - \frac{g}{\|g\|_2}\Big\|_2 \le \frac{2B}{\log(1+2\eta)\;\|g\|_2\;t} = \mathcal{O}(1/t).$$

The rate is tight in general: $r_t \to r_\infty = \theta_0 + Z^\dagger v$, and whenever r_∞ has a component orthogonal to g the direction error is of exact order $1/t$. Finally, $Z\theta = \mathbf{1}$ is equivalent to $\langle \theta, x_i\rangle = y_i$ because $y_i^2 = 1$, and $g = Z^\top K^{-1}\mathbf{1} = X^\top(XX^\top)^{-1}y = X^\dagger y$ is its minimum-norm solution.

(4) Margin. Since $Zg = \mathbf{1}$, all logits of g equal 1, so $\min_i \langle g/\|g\|_2, z_i\rangle = 1/\|g\|_2$, and $\|g\|_2^2 = \mathbf{1}^\top K^{-1} ZZ^\top K^{-1}\mathbf{1} = \mathbf{1}^\top K^{-1}\mathbf{1}$. By (3) and continuity of $u \mapsto \min_i \langle u, z_i\rangle$, the limiting margin is $(\mathbf{1}^\top K^{-1}\mathbf{1})^{-1/2}$. The hard-margin SVM is the convex program $\min\{\frac{1}{2}\|\theta\|_2^2 : Z\theta \ge \mathbf{1}\}$, with unique solution θ_{SVM} and $\gamma_2 = 1/\|\theta_{\mathrm{SVM}}\|_2$. Its constraints are linear, so its KKT conditions are necessary and sufficient:

$$\theta = Z^\top\alpha, \qquad \alpha \ge 0, \qquad Z\theta \ge \mathbf{1}, \qquad \alpha_i\big(\langle\theta, z_i\rangle - 1\big) = 0 \;\; \forall i.$$

Inequality. g is feasible, so $\|\theta_{\mathrm{SVM}}\|_2 \le \|g\|_2$, i.e. $(\mathbf{1}^\top K^{-1}\mathbf{1})^{-1/2} \le \gamma_2$. Equality holds if and only if g attains the minimum, i.e. if and only if $g = \theta_{\mathrm{SVM}}$. *If* $K^{-1}\mathbf{1} \ge 0$, *then* $g = \theta_{\mathrm{SVM}}$. Take $\alpha := K^{-1}\mathbf{1}$. Then $g = Z^\top\alpha$, $\alpha \ge 0$ by assumption, $Zg = \mathbf{1}$, and every complementary-slackness product vanishes. By sufficiency of the KKT conditions, g is optimal. *If* $g = \theta_{\mathrm{SVM}}$, *then* $K^{-1}\mathbf{1} \ge 0$. By necessity there is $\alpha \ge 0$ with $g = Z^\top\alpha$. Multiplying by Z gives $Zg = K\alpha$, and $Zg = \mathbf{1}$, so $\alpha = K^{-1}\mathbf{1}$; hence $K^{-1}\mathbf{1} \ge 0$. Here $\mathrm{rank}Z = n$ is used twice: K is invertible, and the representation $g = Z^\top\alpha$ is unique. *Interpretation.* Call z_i a support vector of the SVM if it lies on the margin, $\langle\theta_{\mathrm{SVM}}, z_i\rangle = 1$. If every sample is a support vector, then $Z\theta_{\mathrm{SVM}} = \mathbf{1}$, so θ_{SVM} belongs to $\{\theta : Z\theta = \mathbf{1}\} \subset \{\theta : Z\theta \ge \mathbf{1}\}$; it minimizes the norm over the larger set, hence over the smaller one, whose unique minimum-norm element is g, so $g = \theta_{\mathrm{SVM}}$. Conversely, $g = \theta_{\mathrm{SVM}}$ gives $Z\theta_{\mathrm{SVM}} = Zg = \mathbf{1}$. Together with the two directions above,

$$K^{-1}\mathbf{1} \ge 0 \iff g = \theta_{\mathrm{SVM}} \iff \text{every sample lies on the margin of the SVM},$$

which is the interpretation in the statement. This geometric notion of support vector is weaker than asking for a positive multiplier: for $z_1 = (1,0)$ and $z_2 = (1,1)$ one has $K^{-1}\mathbf{1} = (1,0)$, both samples lie on the margin of $\theta_{\mathrm{SVM}} = (1,0)$, but $\alpha_2 = 0$. □

Proposition F.2 (Squared prediction distance)**.** *Let* $D_{\mathrm{LS}}(\theta,\theta') = \frac{1}{2n}\sum_i(\langle\theta,x_i\rangle - \langle\theta',x_i\rangle)^2 = \frac{1}{2n}\|Z(\theta-\theta')\|_2^2$, *the Bregman divergence of* $w_{\mathrm{LS}}(\theta) = \frac{1}{2}\theta^\top\widehat{\Sigma}\theta$ *with* $\widehat{\Sigma} = Z^\top Z/n$. *(a) If* $\mathrm{rank}Z = d$, *then* $\widehat{\Sigma} \succ 0$ *and Corollary 3.7 applies with* $A = \widehat{\Sigma}/2$*: BPPA converges in direction to the maximum* $\|\cdot\|_{\widehat{\Sigma}^{-1}}$*-margin classifier. Up to the factor* $1/2$*, which does not affect the direction, this is the divergence* $D^{(2)}$ *of Appendix G. (b) If* $\mathrm{rank}Z = n \le d$, *then under the selection of Lemma 5.2 and the exponential loss the logits follow the common recursion* $s' = s + 2\eta e^{-s'}$, *the differences* $s_{t,i} - s_{t,j}$ *tend to* 0, $s_{t,i} = \log(2\eta t)(1+o(1))$, $\mathcal{L}(\theta_t) = \Theta(1/t)$, *and* $\theta_t/\|\theta_t\|_2 \to Z^\dagger\mathbf{1}/\left\|Z^\dagger\mathbf{1}\right\|_2$ *at rate* $\mathcal{O}(1/\log t)$.

Proof. (a) is immediate. For (b), the per-sample first-order condition is $-e^{-s'} + \frac{1}{2\eta}(s'-s) = 0$. The map $f : s \mapsto s'$ is increasing with $f'(s) = 1/(1+2\eta e^{-f(s)}) \in (0,1)$ (implicit differentiation). For two logits $a_t \ge b_t$ following the same map, the mean value theorem gives $a_{t+1} - b_{t+1} \le (a_t - b_t)/(1+2\eta e^{-a_{t+1}})$. Since $2\eta e^{-a_{t+1}} = a_{t+1} - a_t$ and $\log(1+x) \ge x/(1+2\eta)$ for $0 \le x \le 2\eta$ (valid once $a_t \ge 0$), $\prod_{k<t}(1+2\eta e^{-a_{k+1}})^{-1} \le \exp(-(a_t - a_0)/(1+2\eta))$, which tends to 0 because $a_t \to \infty$: if (a_t) were bounded, its increments $2\eta e^{-a_{t+1}}$ would be bounded below, a contradiction. The asymptotics $e^{a_{t+1}} - e^{a_t} = e^{a_{t+1}}(1 - e^{-(a_{t+1}-a_t)}) \to 2\eta$ give $e^{a_t} \sim 2\eta t$, hence $s_{t,i} = \log(2\eta t)(1+o(1))$ and $\mathcal{L}(\theta_t) = \Theta(1/t)$. Finally $\theta_t = \theta_0 + Z^\dagger(s_t - s_0)$ with $s_t = \bar{s}_t\mathbf{1} + o(1)$ and $\bar{s}_t \to \infty$, so the direction converges to $Z^\dagger\mathbf{1}$ with error $\mathcal{O}(1/\bar{s}_t) = \mathcal{O}(1/\log t)$. □

G Details of the Numerical Illustrations

Numerical methods. Each Bregman proximal step is solved by a damped Newton method started at the previous iterate, with the optimality residual $\nabla w(\theta_{t+1}) - \nabla w(\theta_t) + 2\eta_t\nabla\mathcal{L}(\theta_{t+1})$ of order 10^{-15}. Maximum-margin reference values are computed exactly: the ℓ_2-SVM by enumerating the KKT active sets, and the ℓ_1-maximum margin by linear programming followed by an exact verification in rational arithmetic.

Constant and adaptive stepsizes (Figure 2)

This experiment checks Proposition 3.6, the results for strongly convex and smooth DGFs. The dataset $\mathcal{S}$ has the positive samples $(-0.5, 1)$ and $(2, 1)$ and the negative samples $(-0.5, -1)$ and $(-0.75, -1)$; its maximum ℓ_2-margin classifier is $u_2 = (0, 1)$. The DGF is $w(\theta) = \|\theta\|_2^2$, so that $D_w(x,y) = \|x-y\|_2^2$, $\mu_w = 2$ and ${R_*}^2 = \max_i\|x_i\|_2^2 = 5$; both methods start at the origin and run for 1200 iterations. Constant-stepsize BPPA uses $\eta = 1$, and constant-stepsize MD uses $\eta = \mu_w/(2{R_*}^2) = 0.2$, the largest stepsize covered by Proposition 3.6(3). The adaptive schedule is $\eta_t = 1/(\mathcal{L}(\theta_t)\sqrt{t+1})$ for BPPA and $\eta_t = \min\{0.2, 1/\sqrt{t+1}\}/\mathcal{L}(\theta_t)$ for MD. All four runs converge in direction to u_2, as Corollaries 3.7 and D.3 and Proposition 3.6 with $\kappa_w = 1$ guarantee, and the adaptive schedules are much faster. The loss is monotone in all four runs. After 1200

iterations: BPPA, loss 8.3×10^{-4} and cosine similarity 0.9884 with $\eta = 1$, loss 3.5×10^{-27} and cosine similarity 0.9998 with adaptive steps; MD, loss 4.1×10^{-3} and cosine similarity 0.9816 with $\eta = 0.2$, loss 9.4×10^{-29} and cosine similarity 0.9999 with adaptive steps.

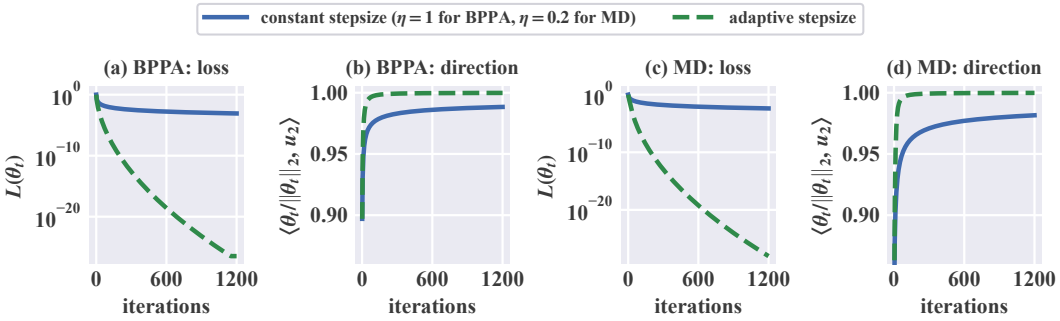

Figure 2: Proposition 3.6: BPPA and MD on $\mathcal{S}$ with $D_w(x, y) = \|x - y\|_2^2$. Loss (a, c) and cosine similarity with the maximum-margin direction u_2 (b, d).

Sharpness of the alignment constant (Figure 1(a) and Figure 3)

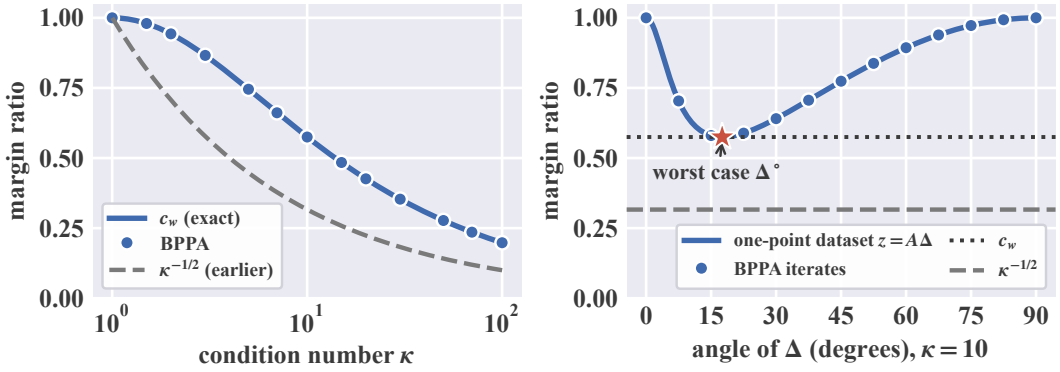

Figure 3: Theorem 3.8. (a) Margin ratio of BPPA on the worst-case one-point dataset as a function of the condition number κ. (b) Margin ratio on one-point datasets $z = A\Delta$ as a function of the direction of Δ, for $\kappa = 10$; the star marks Δ°.

Panel (a): $A = \mathrm{diag}(1, \kappa)$ for thirteen values of $\kappa \in [1, 100]$, the Kantorovich direction $\Delta^\circ \propto (1, \kappa^{-1/2})$ and the one-point dataset $z = A\Delta^\circ$. BPPA with $w(\theta) = \frac{1}{2}\theta^\top A\theta$, $\eta = 1$ and $\theta_0 = 0$ runs for 300 steps; each step is the scalar equation $\theta_{t+1} = \theta_t + \rho\,\Delta^\circ$, $\rho = 2\eta e^{-\langle \theta_t + \rho\Delta^\circ, z\rangle}$, solved by bisection. The largest deviation of the ℓ_2-margin ratio from $2\sqrt{\kappa}/(1+\kappa)$ over all steps and all κ is 3.3×10^{-16}. Panel (b): $\kappa = 10$, $\Delta = (\cos\varphi, \sin\varphi)$ for thirteen angles φ and 100 steps each; the curve is $\langle A\Delta, \Delta\rangle / (\|\Delta\|_2 \|A\Delta\|_2)$, and Δ° corresponds to $\varphi = \arctan(\kappa^{-1/2}) \approx 17.5°$.

Parameter-space KL (Figure 1(b, c) and Figures 4–5)

The signed samples are the rows of

$$Z = \begin{pmatrix} 0.1 & 0.3 & 0.4 & 0.9 & -0.1 \\ -0.3 & 0.9 & 0.7 & 0.7 & 0.2 \\ 0 & 0.9 & -0.4 & 0.4 & 0.1 \end{pmatrix}.$$

The pair $u^* = (0, \frac{5}{11}, 0, \frac{6}{11}, 0)$ and $q^* = (\frac{5}{11}, 0, \frac{6}{11})$ proves $\gamma_1 = \frac{69}{110}$ exactly: $\|u^*\|_1 = 1$ and $Zu^* = (\frac{69}{110}, \frac{87}{110}, \frac{69}{110})$, so the primal value is $\frac{69}{110}$; $q^* \in \Delta_3$ and $Z^\top q^* = \frac{1}{110}(5, 69, -4, 69, 1)$, so the dual value $\|Z^\top q^*\|_\infty$ is also $\frac{69}{110}$. Every maximizer u satisfies $\langle u, Z^\top q^*\rangle = \frac{69}{110}$, which forces $u_j = 0$ for

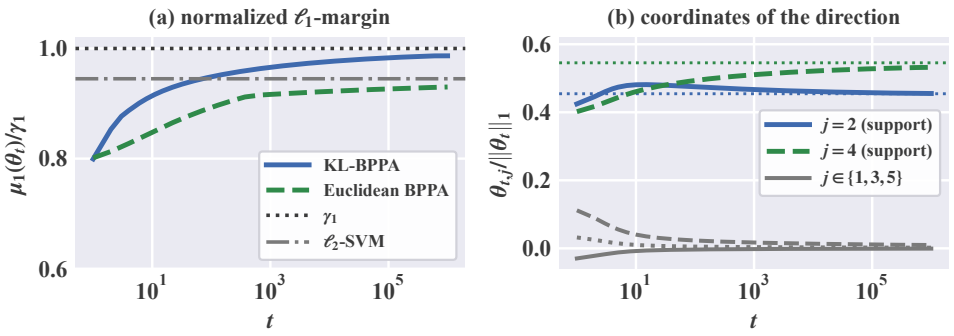

Figure 4: Theorem 4.2 on a dataset with a unique sparse maximizer u^*, $\eta = 1$, $\beta = 0.1 \cdot \mathbf{1}$. (a) Normalized ℓ_1-margin of KL-BPPA and of Euclidean BPPA. (b) Coordinates of $\theta_t / \|\theta_t\|_1$; dotted lines mark u_2^* and u_4^*.

$j \in \{1,3,5\}$, where $|(Z^\top q^*)_j| < \frac{69}{110}$, and $\langle u, z_1\rangle = \langle u, z_3\rangle = \frac{69}{110}$ on the remaining coordinates; this system has the unique solution u^*. The strict-complementarity gap off the support is $\frac{64}{110}$. KL-BPPA uses $\beta = 0.1 \cdot \mathbf{1}$, $\eta = 1$ and 10^6 steps; Euclidean BPPA uses $w(\theta) = \frac{1}{2}\|\theta\|_2^2$ and $\eta = 1$. The ℓ_1-margin of the ℓ_2-SVM direction is $0.945\,\gamma_1$. After 10^6 steps, KL-BPPA has $\mu_1(\theta_t) = 0.987\,\gamma_1$, direction error 0.027 and normalized coordinates $(-0.0008, 0.4553, 0.0098, 0.5321, 0.0020)$, and Euclidean BPPA has $\mu_1(\theta_t) = 0.930\,\gamma_1$.

Figure 5 shows the margin gap and the ℓ_1-distance of the direction to u^*; both decrease slowly, consistently with the absence of a rate in Theorem 4.2.

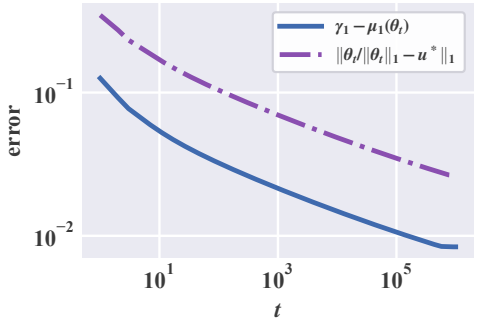

Figure 5: Theorem 4.2: margin gap and ℓ_1-distance of $\theta_t / \|\theta_t\|_1$ to u^* for KL-BPPA on the dataset of Figure 4.

Prediction-space KL (Figure 1(d) and Figure 6)

One proximal step lowers the margin from $\gamma_2 = 1$ to 0.012 (KL) and 0.013 (LS). The KL direction error is of exact order $1/t$, with t times the error converging to 1.375×10^{-3}; in this example the LS direction converges faster, so the $\mathcal{O}(1/\log t)$ of Proposition F.2 is only an upper bound.

The samples are $z_1 = (1, 0)$ and $z_2 = (10, 0.1)$, so $K^{-1}\mathbf{1} = (9001, -900)$, $Z^\dagger \mathbf{1} = (1, -90)$ and the limiting margin is $1/\sqrt{8101} = 0.0111104$; $\theta_0 = (0.2, 0)$ is the SVM direction and $\eta = 1$. The logits follow the scalar maps of Lemma F.1 for KL and of Proposition F.2 for LS. The KL map is evaluated in the form $u = u'[1 + 2\eta(1+u)(1+u')]$ with $u = e^{-s}$, which avoids subtracting two sigmoids close to 1, and the excess increments $e_{k,i}$ are computed with $\log(1+x)$ evaluated directly for small x, so that $\varepsilon_{t,i} = \sum_{k \geq t} e_{k,i}$ is accurate far below machine precision relative to the logits. The parameters are $\theta_t = \theta_0 + Z^\dagger(s_t - s_0)$. The KL direction errors at $t = 10^2, 10^3, 10^4, 2 \times 10^4$ are 1.365×10^{-5}, 1.374×10^{-6}, 1.375×10^{-7} and 6.875×10^{-8}; the LS ones are 8.93×10^{-6}, 6.25×10^{-7}, 4.79×10^{-8} and 2.24×10^{-8}.

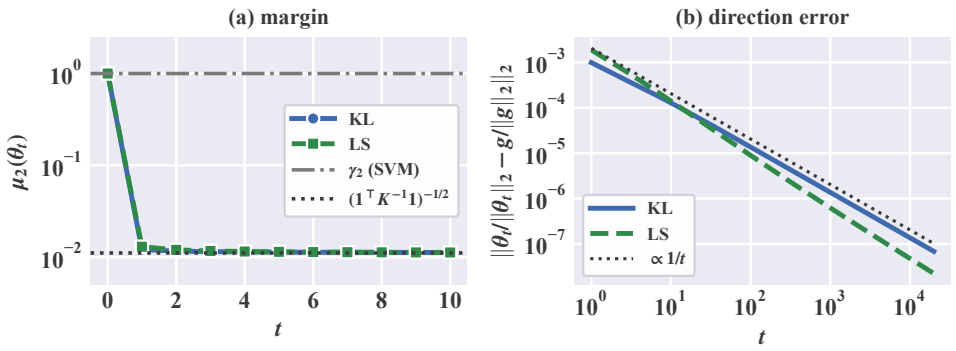

Figure 6: Theorem 5.3 on the two-point example, KL versus LS (squared prediction distance) proximal steps, $\eta = 1$. (a) Normalized ℓ_2-margin. (b) Distance of the direction to $g = Z^\dagger \mathbf{1}$.

H CIFAR-100 Experiments

We apply BPPA to CIFAR-100 (Krizhevsky, 2009) with ResNet-18 (He et al., 2016), ShuffleNetV2 (Ma et al., 2018) and MobileNetV2 (Sandler et al., 2018), with two divergences widely used in practice, the squared prediction distance (Tarvainen & Valpola, 2017) and the KL divergence between predictions (Furlanello et al., 2018),

$$D_{\mathrm{LS}}(\theta', \theta) = \frac{1}{2n} \sum_{i=1}^{n} \|f_\theta(x_i) - f_{\theta'}(x_i)\|_2^2, \qquad D_{\mathrm{KL}}(\theta, \theta') = \frac{1}{2n} \sum_{i=1}^{n} \mathrm{KL}\big(f_{\theta'}(x_i) \,\|\, f_\theta(x_i)\big).$$

Both are defined on the predictions of the network f_θ.

Each BPPA iteration solves

$$\theta_{t+1} = \operatorname*{argmin}_\theta \frac{1}{n} \sum_{i=1}^{n} \ell\big(f_\theta(x_i); y_i\big) + \frac{1}{2\eta_t} D(\theta; \theta_t),$$

where D is D_{KL} or D_{LS} and θ_0 is obtained by standard SGD training. Each proximal step is solved approximately by SGD with batch size 128 and an initial learning rate of 0.1, divided by 5 at epochs 60, 120 and 160, and standard SGD training uses the same batch size and schedule. We run 3 proximal steps with $\eta = 0.025$ for D_{KL} and $\eta = 0.2$ for D_{LS}. Table 3 lists the number of epochs.

Method	Prox. steps	Ep. / step	Total ep.
SGD baseline	—	—	200
BPPA (D_{KL}/D_{LS})	3	200	800

Table 3: Epochs used in the CIFAR-100 experiments. BPPA uses four times the SGD epochs.

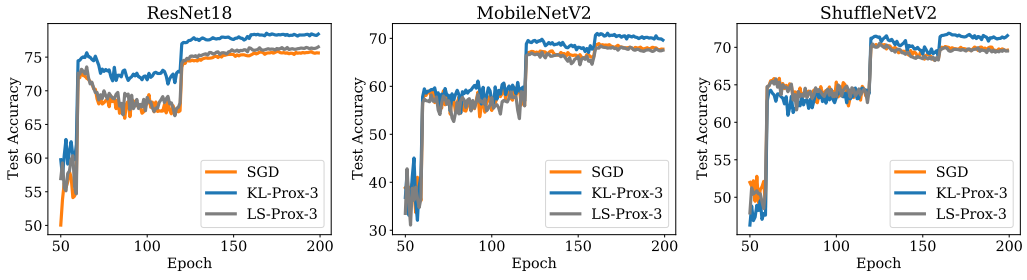

Figure 7: CIFAR-100: test accuracy of SGD and of BPPA with D_{KL} and D_{LS}, for ResNet-18, MobileNetV2 and ShuffleNetV2.

Figure 7 shows that, across architectures, BPPA with D_{KL} outperforms standard training with SGD, while BPPA with D_{LS} yields negligible differences compared with SGD. For ResNet-18, the test

accuracy increases from 75.83% with standard training to 78.56% after three proximal steps with D_{KL}. The qualitative difference between D_{KL} and D_{LS} suggests that the divergence plays an important role in the performance of the model learned by BPPA.