The Fast Convergence of Boosting

Matus Telgarsky
Department of Computer Science and Engineering
University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093-0404
mtelgars@cs.ucsd.edu

Abstract

This manuscript considers the convergence rate of boosting under a large class of losses, including the exponential and logistic losses, where the best previous rate of convergence was $\mathcal{O}(\exp(1/\epsilon^2))$. First, it is established that the setting of weak learnability aids the entire class, granting a rate $\mathcal{O}(\ln(1/\epsilon))$. Next, the (disjoint) conditions under which the infimal empirical risk is attainable are characterized in terms of the sample and weak learning class, and a new proof is given for the known rate $\mathcal{O}(\ln(1/\epsilon))$. Finally, it is established that any instance can be decomposed into two smaller instances resembling the two preceding special cases, yielding a rate $\mathcal{O}(1/\epsilon)$, with a matching lower bound for the logistic loss. The principal technical hurdle throughout this work is the potential unattainability of the infimal empirical risk; the technique for overcoming this barrier may be of general interest.

1 Introduction

Boosting is the task of converting inaccurate *weak learners* into a single accurate predictor. The existence of any such method was unknown until the breakthrough result of Schapire [1]: under a *weak learning assumption*, it is possible to combine many carefully chosen weak learners into a majority of majorities with arbitrarily low training error. Soon after, Freund [2] noted that a single majority is enough, and that $\mathcal{O}(\ln(1/\epsilon))$ iterations are both necessary and sufficient to attain accuracy ϵ. Finally, their combined effort produced AdaBoost, which attains the optimal convergence rate (under the weak learning assumption), and has an astonishingly simple implementation [3].

It was eventually revealed that AdaBoost was minimizing a risk functional, specifically the exponential loss [4]. Aiming to alleviate perceived deficiencies in the algorithm, other loss functions were proposed, foremost amongst these being the logistic loss [5]. Given the wide practical success of boosting with the logistic loss, it is perhaps surprising that no convergence rate better than $\mathcal{O}(\exp(1/\epsilon^2))$ was known, even under the weak learning assumption [6]. The reason for this deficiency is simple: unlike SVM, least squares, and basically any other optimization problem considered in machine learning, there might not exist a choice which attains the minimal risk! This reliance is carried over from convex optimization, where the assumption of attainability is generally made, either directly, or through stronger conditions like compact level sets or strong convexity [7].

Convergence rate analysis provides a valuable mechanism to compare and improve of minimization algorithms. But there is a deeper significance with boosting: a convergence rate of $\mathcal{O}(\ln(1/\epsilon))$ means that, with a combination of just $\mathcal{O}(\ln(1/\epsilon))$ predictors, one can construct an ϵ-optimal classifier, which is crucial to both the computational efficiency and statistical stability of this predictor.

The contribution of this manuscript is to provide a tight convergence theory for a large class of losses, including the exponential and logistic losses, which has heretofore resisted analysis. The goal is a general analysis without any assumptions (attainability of the minimum, or weak learnability),

1

however this manuscript also demonstrates how the classically understood scenarios of attainability and weak learnability can be understood directly from the sample and the weak learning class.

The organization is as follows. Section 2 provides a few pieces of background: how to encode the weak learning class and sample as a matrix, boosting as coordinate descent, and the primal objective function. Section 3 then gives the dual problem, max entropy. Given these tools, section 4 shows how to adjust the weak learning rate to a quantity which is useful without any assumptions. The first step towards convergence rates is then taken in section 5, which demonstrates that the weak learning rate is in fact a mechanism to convert between the primal and dual problems.

The convergence rates then follow: section 6 and section 7 discuss, respectively, the conditions under which classical weak learnability and (disjointly) attainability hold, both yielding the rate $\mathcal{O}(\ln(1/\epsilon))$, and finally section 8 shows how the general case may be decomposed into these two, and the conflicting optimization behavior leads to a degraded rate of $\mathcal{O}(1/\epsilon)$. The last section will also exhibit an $\Omega(1/\epsilon)$ lower bound for the logistic loss.

1.1 Related Work

The development of general convergence rates has a number of important milestones in the past decade. The first convergence result, albeit without any rates, is due to Collins et al. [8]; the work considered the improvement due to a single step, and as its update rule was less aggressive than the line search of boosting, it appears to imply general convergence. Next, Bickel et al. [6] showed a rate of $\mathcal{O}(\exp(1/\epsilon^2))$, where the assumptions of bounded second derivatives on compact sets are also necessary here.

Many extremely important cases have also been handled. The first is the original rate of $\mathcal{O}(\ln(1/\epsilon))$ for the exponential loss under the weak learning assumption [3]. Next, Rätsch et al. [9] showed, for a class of losses similar to those considered here, a rate of $\mathcal{O}(\ln(1/\epsilon))$ when the loss minimizer is attainable. The current manuscript provides another mechanism to analyze this case (with the same rate), which is crucial to being able to produce a general analysis. And, very recently, parallel to this work, Mukherjee et al. [10] established the general convergence under the exponential loss, with a rate of $\Theta(1/\epsilon)$. The same matrix, due to Schapire [11], was used to show the lower bound there as for the logistic loss here; their upper bound proof also utilized a decomposition theorem.

It is interesting to mention that, for many variants of boosting, general convergence rates were known. Specifically, once it was revealed that boosting is trying to be not only correct but also have large margins [12], much work was invested into methods which explicitly maximized the margin [13], or penalized variants focused on the inseparable case [14, 15]. These methods generally impose some form of regularization [15], which grants attainability of the risk minimizer, and allows standard techniques to grant general convergence rates. Interestingly, the guarantees in those works cited in this paragraph are $\mathcal{O}(1/\epsilon^2)$.

2 Setup

A view of boosting, which pervades this manuscript, is that the action of the weak learning class upon the sample can be encoded as a matrix [9, 15]. Let a sample $\mathcal{S} := \{(x_i, y_i)\}_1^m \subseteq (\mathcal{X} \times \mathcal{Y})^m$ and a weak learning class $\mathcal{H}$ be given. For every $h \in \mathcal{H}$, let $\mathcal{S}|_h$ denote the projection onto $\mathcal{S}$ induced by h; that is, $\mathcal{S}|_h$ is a vector of length m, with coordinates $(\mathcal{S}|_h)_i = y_i h(x_i)$. If the set of all such columns $\{\mathcal{S}|_h : h \in \mathcal{H}\}$ is finite, collect them into the matrix $A \in \mathbb{R}^{m \times n}$. Let a_i denote the ith row of A, corresponding to the example (x_i, y_i), and let $\{h_j\}_1^n$ index the set of weak learners corresponding to columns of A. It is assumed, for convenience, that entries of A are within $[-1, +1]$; relaxing this assumption merely scales the presented rates by a constant.

The setting considered in this manuscript is that this finite matrix can be constructed. Note that this can encode infinite classes, so long as they map to only $k < \infty$ values (in which case A has at most k^m columns). As another example, if the weak learners are binary, and $\mathcal{H}$ has VC dimension d, then Sauer's lemma grants that A has at most $(m + 1)^d$ columns. This matrix view of boosting is thus similar to the interpretation of boosting performing descent on functional space, but the class complexity and finite sample have been used to reduce the function class to a finite object [16, 5].

```
┌─────────────────────────────────────────────────────────────────┐
│  Routine BOOST.                                                   │
│  Input Convex function $f \circ A$.                               │
│  Output Approximate primal optimum $\lambda$.                     │
│      1. Initialize $\lambda_0 := \mathbf{0}_n$.                    │
│      2. For $t = 1, 2, \ldots$, while $\nabla(f \circ A)(\lambda_{t-1}) \neq \mathbf{0}_n$: │
│          (a) Choose column $j_t := \operatorname{argmax}_j |\nabla(f \circ A)(\lambda_{t-1})^\top \mathbf{e}_j|$. │
│          (b) Line search: $\alpha_t$ apx. minimizes $\alpha \mapsto (f \circ A)(\lambda_{t-1} + \alpha \mathbf{e}_{j_t})$. │
│          (c) Update $\lambda_t := \lambda_{t-1} + \alpha_t \mathbf{e}_{j_t}$. │
│      3. Return $\lambda_{t-1}$.                                    │
└─────────────────────────────────────────────────────────────────┘
```

Figure 1: l^1 steepest descent [17, Algorithm 9.4] applied to $f \circ A$.

To make the connection to boosting, the missing ingredient is the loss function. Let $\mathbb{G}_0$ denote the set of loss functions g satisfying: g is twice continuously differentiable, $g'' > 0$ (which implies strict convexity), and $\lim_{x \to \infty} g(x) = 0$. (A few more conditions will be added in section 5 to prove convergence rates, but these properties suffice for the current exposition.) Crucially, the exponential loss $\exp(-x)$ from AdaBoost and the logistic loss $\ln(1 + \exp(-x))$ are in $\mathbb{G}_0$ (and the eventual $\mathbb{G}$).

Boosting determines some weighting $\lambda \in \mathbb{R}^n$ of the columns of A, which correspond to weak learners in $\mathcal{H}$. The (unnormalized) margin of example i is thus $\langle a_i, \lambda \rangle = \mathbf{e}_i^\top A \lambda$, where $\mathbf{e}_i$ is an indicator vector. Since the prediction on x_i is $\mathbb{1}[\langle a_i, \lambda \rangle \geq 0]$, it follows that $A\lambda > \mathbf{0}_m$ (where $\mathbf{0}_m$ is the zero vector) implies a training error of zero. As such, boosting solves the minimization problem

$$\inf_{\lambda \in \mathbb{R}^n} \sum_{i=1}^m g(\langle a_i, \lambda \rangle) = \inf_{\lambda \in \mathbb{R}^n} \sum_{i=1}^m g(\mathbf{e}_i^\top A \lambda) = \inf_{\lambda \in \mathbb{R}^n} f(A\lambda) = \inf_{\lambda \in \mathbb{R}^n} (f \circ A)(\lambda) =: \bar{f}_A, \quad (2.1)$$

where $f : \mathbb{R}^m \to \mathbb{R}$ is the convenience function $f(x) = \sum_i g((x)_i)$, and in the present problem denotes the (unnormalized) empirical risk. $\bar{f}_A$ will denote the optimal objective value.

The infimum in eq. (2.1) may well not be attainable. Suppose there exists λ' such that $A\lambda' > \mathbf{0}_m$ (theorem 6.1 will show that this is equivalent to the weak learning assumption). Then

$$0 \leq \inf_{\lambda \in \mathbb{R}^n} f(A\lambda) \leq \inf \{f(A\lambda) : \lambda = c\lambda', c > 0\} = \inf_{c > 0} f(c(A\lambda')) = 0.$$

On the other hand, for any $\lambda \in \mathbb{R}^n$, $f(A\lambda) > 0$. Thus the infimum is never attainable when weak learnability holds.

The template boosting algorithm appears in fig. 1, formulated in terms of $f \circ A$ to make the connection to coordinate descent as clear as possible. To interpret the gradient terms, note that

$$(\nabla(f \circ A)(\lambda))_j = (A^\top \nabla f(A\lambda))_j = \sum_{i=1}^m g'(\langle a_i, \lambda \rangle) h_j(x_i) y_i,$$

which is the expected correlation of h_j with the target labels according to an unnormalized distribution with weights $g'(\langle a_i, \lambda \rangle)$. The stopping condition $\nabla(f \circ A)(\lambda) = \mathbf{0}_m$ means: either the distribution is degenerate (it is exactly zero), or every weak learner is uncorrelated with the target.

As such, eq. (2.1) represents an equivalent formulation of boosting, with one minor modification: the column (weak learner) selection has an absolute value. But note that this is the same as closing $\mathcal{H}$ under complementation (i.e., for any $h \in \mathcal{H}$, there exists h' with $h(x) = -h'(x)$), which is assumed in many theoretical treatments of boosting.

In the case of the exponential loss with binary weak learners, the line search step has a convenient closed form; but for other losses, or even for the exponential loss but with confidence-rated predictors, there may not be a closed form. Moreover, this univariate search problem may lack a minimizer. To produce the eventual convergence rates, this manuscript utilizes a step size minimizing an upper bounding quadratic (which is guaranteed to exist); if instead a standard iterative line search guarantee were used, rates would only degrade by a constant factor [17, section 9.3.1].

As a final remark, consider the rows $\{a_i\}_1^m$ of A as a collection of m points in $\mathbb{R}^n$. Due to the form of g, BOOST is therefore searching for a halfspace, parameterized by a vector λ, which contains all of the points. Sometimes such a halfspace may not exist, and g applies a smoothly increasing penalty to points that are farther and farther outside it.

3 Dual Problem

This section provides the convex dual to eq. (2.1). The relevance of the dual to convergence rates is as follows. First, although the primal optimum may not be attainable, the dual optimum is always attainable—this suggests a strategy of mapping the convergence strategy to the dual, where there exists a clear notion of progress to the optimum. Second, this section determines the dual feasible set—the space of dual variables or what the boosting literature typically calls *unnormalized weights*. Understanding this set is key to relating weak learnability, attainability, and general instances.

Before proceeding, note that the dual formulation will make use of the Fenchel conjugate $h^*(\phi) = \sup_{x \in \mathrm{dom}(h)} \langle x, \phi \rangle - h(x)$, a concept taking a central place in convex analysis [18, 19]. Interestingly, the Fenchel conjugates to the exponential and logistic losses are respectively the Boltzmann-Shannon and Fermi-Dirac entropies [19, Commentary, section 3.3], and thus the dual is explicitly performing entropy maximization (cf. lemma C.2). As a final piece of notation, denote the kernel of a matrix $B \in \mathbb{R}^{m \times n}$ by $\mathrm{Ker}(B) = \{\phi \in \mathbb{R}^n : B\phi = \mathbf{0}_m\}$.

Theorem 3.1. *For any $A \in \mathbb{R}^{m \times n}$ and $g \in \mathbb{G}_0$ with $f(x) = \sum_i g((x)_i)$,*

$$\inf \{f(A\lambda) : \lambda \in \mathbb{R}^n\} = \sup \{-f^*(-\phi) : \phi \in \Phi_A\}, \tag{3.2}$$

where $\Phi_A := \mathrm{Ker}(A^\top) \cap \mathbb{R}_+^m$ is the dual feasible set. The dual optimum ψ_A is unique and attainable. Lastly, $f^*(\phi) = \sum_{i=1}^m g^*((\phi)_i)$.

The dual feasible set $\Phi_A = \mathrm{Ker}(A^\top) \cap \mathbb{R}_+^m$ has a strong interpretation. Suppose $\psi \in \Phi_A$; then ψ is a nonnegative vector (since $\psi \in \mathbb{R}_+^m$), and, for any j, $0 = (\phi^\top A)_j = \sum_{i=1}^m \phi_i y_i h_j(x_i)$. That is to say, every nonzero feasible dual vector provides a (an unnormalized) distribution upon which every weak learner is uncorrelated! Furthermore, recall that the weak learning assumption states that under any weighting of the input, there exists a correlated weak learner; as such, weak learnability necessitates that the dual feasible set contains only the zero vector.

There is also a geometric interpretation. Ignoring the constraint, $-f^*$ attains its maximum at some rescaling of the uniform distribution (for details, please see lemma C.2). As such, the constrained dual problem is aiming to write the origin as a high entropy convex combination of the points $\{a_i\}_1^m$.

4 A Generalized Weak Learning Rate

The weak learning rate was critical to the original convergence analysis of AdaBoost, providing a handle on the progress of the algorithm. Recall that the quantity appeared in the denominator of the convergence rate, and a *weak learning assumption* critically provided that this quantity is nonzero. This section will generalize the weak learning rate to a quantity which is always positive, without any assumptions.

Note briefly that this manuscript will differ slightly from the norm in that weak learning will be a purely *sample-specific* concept. That is, the concern here is convergence, and all that matters is the sample $\mathcal{S} = \{(x_i, y_i)\}_1^m$, as encoded in A; it doesn't matter if there are wild points outside this sample, because the algorithm has no access to them.

This distinction has the following implication. The usual weak learning assumption states that there exists no uncorrelating distribution over the input *space*. This of course implies that any training sample $\mathcal{S}$ used by the algorithm will also have this property; however, it suffices that there is no distribution over the input *sample* $\mathcal{S}$ which uncorrelates the weak learners from the target.

Returning to task, the weak learning assumption posits the existence of a constant, the weak learning rate γ, which lower bounds the correlation of the best weak learner with the target for any distribu-

tion. Stated in terms of the matrix A,

$$0 < \gamma = \inf_{\substack{\phi \in \mathbb{R}^m_+ \\ \|\phi\|=1}} \max_{j \in [n]} \left| \sum_{i=1}^m (\phi)_i y_i h_j(x_i) \right| = \inf_{\phi \in \mathbb{R}^m_+ \setminus \{\mathbf{0}_m\}} \frac{\|A^\top \phi\|_\infty}{\|\phi\|_1} = \inf_{\phi \in \mathbb{R}^m_+ \setminus \{\mathbf{0}_m\}} \frac{\|A^\top \phi\|_\infty}{\|\phi - \mathbf{0}_m\|_1}.$$

$$(4.1)$$

The only way this quantity can be positive is if $\phi \notin \text{Ker}(A^\top) \cap \mathbb{R}^m_+ = \Phi_A$, meaning the dual feasible set is exactly $\{\mathbf{0}_m\}$. As such, one candidate adjustment is to simply replace $\{\mathbf{0}_m\}$ with the dual feasible set:

$$\gamma' := \inf_{\phi \in \mathbb{R}^m_+ \setminus \Phi_A} \frac{\|A^\top \phi\|_\infty}{\inf_{\psi \in \Phi_A} \|\phi - \psi\|_1}.$$

Indeed, by the forthcoming proposition 4.3, $\gamma' > 0$ as desired. Due to technical considerations which will be postponed until the various convergence rates, it is necessary to tighten this definition with another set.

Definition 4.2. For a given matrix $A \in \mathbb{R}^{m \times n}$ and set $S \subseteq \mathbb{R}^m$, define

$$\gamma(A, S) := \inf \left\{ \frac{\|A^\top \phi\|_\infty}{\inf_{\psi \in S \cap \text{Ker}(A^\top)} \|\phi - \psi\|_1} : \phi \in S \setminus \text{Ker}(A^\top) \right\}. \qquad \Diamond$$

Crucially, for the choices of S pertinent here, this quantity is always positive.

Proposition 4.3. *Let $A \neq \mathbf{0}_{m \times n}$ and polyhedron S be given. If $S \cap \text{Ker}(A^\top) \neq \emptyset$ and S has nonempty interior, $\gamma(A, S) \in (0, \infty)$.*

To simplify discussion, the following projection and distance notation will be used in the sequel:

$$\mathsf{P}^p_C(x) \in \operatorname*{Argmin}_{y \in C} \|y - x\|_p, \qquad\qquad \mathsf{D}^p_C(x) = \|x - \mathsf{P}^p_C(x)\|_p,$$

with some arbitrary choice made when the minimizer is not unique.

5   Prelude to Convergence Rates: Three Alternatives

The pieces are in place to finally sketch how the convergence rates may be proved. This section identifies how the weak learning rate $\gamma(A, S)$ can be used to convert the standard gradient guarantees into something which can be used in the presence of no attainable minimum. To close, three basic optimization scenarios are identified, which lead to the following three sections on convergence rates. But first, it is a good time to define the final loss function class.

Definition 5.1. Every $g \in \mathbb{G}$ satisfies the following properties. First, $g \in \mathbb{G}_0$. Next, for any $x \in \mathbb{R}^m$ satisfying $f(x) \leq f(A\lambda_0)$, and for any coordinate $(x)_i$, there exist constants $\eta > 0$ and $\beta > 0$ such that $g''((x)_i) \leq \eta g((x)_i)$ and $g((x)_i) \leq -\beta g'((x)_i)$. $\qquad \Diamond$

The exponential loss is in this class with $\eta = \beta = 1$ since $\exp(\cdot)$ is a fixed point with respect to the differentiation operator. Furthermore, as is verified in remark F.1 of the full version, the logistic loss is also in this class, with $\eta = 2^m/(m \ln(2))$ and $\beta \leq 1 + 2^m$. Intuitively, η and β encode how similar some $g \in \mathbb{G}$ is to the exponential loss, and thus these parameters can degrade radically. However, outside the weak learnability case, the other terms in the bounds here will also incur a penalty of the form e^m for the exponential loss, and there is some evidence that this is unavoidable (see the lower bounds in Mukherjee et al. [10] or the upper bounds in Rätsch et al. [9]).

Next, note how the standard guarantee for coordinate descent methods can lead to guarantees on the progress of the algorithm in terms of dual distances, thanks to $\gamma(A, S)$.

Proposition 5.2. *For any t, $A \neq \mathbf{0}^{m \times n}$, $S \supseteq \{-\nabla f(A\lambda_t)\}$ with $\gamma(A, S) > 0$, and $g \in \mathbb{G}$,*

$$f(A\lambda_{t+1}) - \bar{f}_A \leq f(A\lambda_t) - \bar{f}_A - \frac{\gamma(A, S)^2 \mathsf{D}^1_{S \cap \text{Ker}(A^\top)}(-\nabla f(A\lambda_t))^2}{2\eta f(A\lambda_t)}.$$

Proof. The stopping condition grants $-\nabla f(A\lambda_t) \notin \text{Ker}(A^\top)$. Thus, by definition of $\gamma(A, S)$,

$$\gamma(A, S) = \inf_{\phi \in S \setminus \text{Ker}(A^\top)} \frac{\|A^\top \phi\|_\infty}{\mathsf{D}^1_{S \cap \text{Ker}(A^\top)}(\phi)} \leq \frac{\|A^\top \nabla f(A\lambda_t)\|_\infty}{\mathsf{D}^1_{S \cap \text{Ker}(A^\top)}(-\nabla f(A\lambda_t))}.$$

5

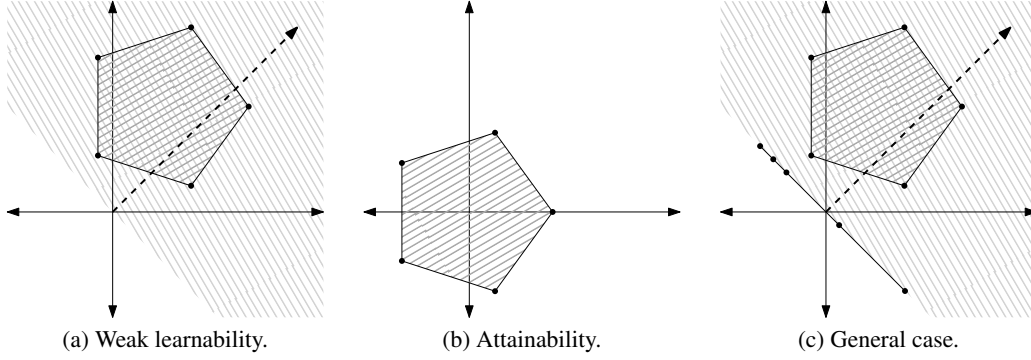|		
(a) Weak learnability.	(b) Attainability.	(c) General case.

Figure 2: Viewing the rows $\{a_i\}_1^m$ of A as points in $\mathbb{R}^n$, boosting seeks a homogeneous halfspace, parameterized by a normal $\lambda \in \mathbb{R}^n$, which contains all m points. The dual, on the other hand, aims to express the origin as a high entropy convex combination of the rows. The convergence rate and dynamics of this process are controlled by A, which dictates one of the three above scenarios.

Combined with a standard guarantee of coordinate descent progress (cf. lemma F.2),

$$f(A\lambda_t) - f(A\lambda_{t+1}) \geq \frac{\|A^\top \nabla f(A\lambda_t)\|_\infty^2}{2\eta f(A\lambda_t)} \geq \frac{\gamma(A,S)^2 \mathsf{D}^1_{S \cap \mathrm{Ker}(A^\top)}(-\nabla f(A\lambda_t))^2}{2\eta f(A\lambda_t)}.$$

Subtracting $\bar{f}_A$ from both sides and rearranging yields the statement. $\qquad \square$

Recall the interpretation of boosting closing section 2: boosting seeks a halfspace, parameterized by $\lambda \in \mathbb{R}^n$, which contains the points $\{a_i\}_1^m$. Progress onward from proposition 5.2 will be divided into three cases, each distinguished by the kind of halfspace which boosting can reach.

These cases appear in fig. 2. The first case is weak learnability: positive margins can be attained on each example, meaning a halfspace exists which strictly contains all points. Boosting races to push all these margins unboundedly large, and has a convergence rate $\mathcal{O}(\ln(1/\epsilon))$. Next is the case that no halfspace contains the points within its interior: either any such halfspace has the points on its boundary, or no such halfspace exists at all (the degenerate choice $\lambda = \mathbf{0}_n$). This is the case of attainability: boosting races towards finite margins at the rate $\mathcal{O}(\ln(1/\epsilon))$.

The final situation is a mix of the two: there exists a halfspace with some points on the boundary, some within its interior. Boosting will try to push some margins to infinity, and keep others finite. These two desires are at odds, and the rate degrades to $\mathcal{O}(1/\epsilon)$. Less metaphorically, the analysis will proceed by decomposing this case into the previous two, applying the above analysis in parallel, and then stitching the result back together. It is precisely while stitching up that an incompatibility arises, and the rate degrades. This is no artifact: a lower bound will be shown for the logistic loss.

6 Convergence Rate under Weak Learnability

To start this section, the following result characterizes weak learnability, including the earlier relationship to the dual feasible set (specifically, that it is precisely the origin), and, as analyzed by many authors, the relationship to separability [1, 9, 15].

Theorem 6.1. *For any $A \in \mathbb{R}^{m \times n}$ and $g \in \mathbb{G}$ the following conditions are equivalent:*

$$\exists \lambda \in \mathbb{R}^n . A\lambda \in \mathbb{R}^m_{++}, \tag{6.2}$$

$$\inf_{\lambda \in \mathbb{R}^n} f(A\lambda) = 0, \tag{6.3}$$

$$\psi_A = \mathbf{0}_m, \tag{6.4}$$

$$\Phi_A = \{\mathbf{0}_m\}. \tag{6.5}$$

The equivalence means the presence of any of these properties suffices to indicate weak learnability. The last two statements encode the usual distributional version of the weak learning assumption.

The first encodes the fact that there exists a homogeneous halfspace containing all points within its interior; this encodes separability, since removing the factor y_i from the definition of a_i will place all negative points outside the halfspace. Lastly, the second statement encodes the fact that the empirical risk approaches zero.

Theorem 6.6. *Suppose $A\lambda > \mathbf{0}_m$ and $g \in \mathbb{G}$; then $\gamma(A, \mathbb{R}_+^m) > 0$, and for all t,*

$$f(A\lambda_t) - \bar{f}_A \leq f(A\lambda_0)\left(1 - \frac{\gamma(A, \mathbb{R}_+^m)^2}{2\beta^2\eta}\right)^t.$$

Proof. By theorem 6.1, $\mathbb{R}_+^m \cap \mathrm{Ker}(A^\top) = \Phi_A = \{\mathbf{0}_m\}$, which combined with $g \leq -\beta g'$ gives

$$\mathrm{D}_{\Phi_A}^1(-\nabla f(A\lambda_t)) = \inf_{\psi \in \Phi_A} \|-\nabla f(A\lambda_t) - \psi\|_1 = \|\nabla f(A\lambda_t)\|_1 \geq f(A\lambda_t)/\beta.$$

Plugging this and $\bar{f}_A = 0$ (again by theorem 6.1) along with polyhedron $\mathbb{R}_+^m \supseteq -\nabla f(\mathbb{R}^m)$ (whereby $\gamma(A, \mathbb{R}_+^m) > 0$ by proposition 4.3 since $\psi_A \in \mathbb{R}_+^m$) into proposition 5.2 gives

$$f(A\lambda_{t+1}) \leq f(A\lambda_t) - \frac{\gamma(A, \mathbb{R}_+^m)^2 f(A\lambda_t)}{2\beta^2\eta} = f(A\lambda_t)\left(1 - \frac{\gamma(A, \mathbb{R}_+^m)^2}{2\beta^2\eta}\right),$$

and recursively applying this inequality yields the result. $\qquad\square$

Since the present setting is weak learnability, note by (4.1) that the choice of polyhedron $\mathbb{R}_+^m$ grants that $\gamma(A, \mathbb{R}_+^m)$ is exactly the original weak learning rate. When specialized for the exponential loss (where $\eta = \beta = 1$), the bound becomes $(1 - \gamma(A, \mathbb{R}_+^m)^2/2)^t$, which exactly recovers the bound of Schapire and Singer [20], although via different analysis.

In general, solving for t in the expression $\epsilon = \frac{f(A\lambda_t) - \bar{f}_A}{f(A\lambda_0) - \bar{f}_A} \leq \left(1 - \frac{\gamma(f,A)^2}{2\beta^2\eta}\right)^t \leq \exp\left(-\frac{\gamma(f,A)^2 t}{2\beta^2\eta}\right)$ reveals that $t \leq \frac{2\beta^2\eta}{\gamma(A,S)^2}\ln(1/\epsilon)$ iterations suffice to reach error ϵ. Recall that β and η, in the case of the logistic loss, have only been bounded by quantities like 2^m. While it is unclear if this analysis of β and η was tight, note that it is plausible that the logistic loss is slower than the exponential loss in this scenario, as it works less in initial phases to correct minor margin violations.

7 Convergence Rate under Attainability

Theorem 7.1. *For any $A \in \mathbb{R}^{m\times n}$ and $g \in \mathbb{G}$, the following conditions are equivalent:*

$$\forall \lambda \in \mathbb{R}^n \,\textbf{.}\, A\lambda \notin \mathbb{R}_+^m \setminus \{\mathbf{0}_m\}, \tag{7.2}$$

$$f \circ A \text{ has minimizers}, \tag{7.3}$$

$$\psi_A \in \mathbb{R}_{++}^m, \tag{7.4}$$

$$\Phi_A \cap \mathbb{R}_{++}^m \neq \emptyset. \tag{7.5}$$

Interestingly, as revealed in (7.4) and (7.5), attainability entails that the dual has fully interior points, and furthermore that the dual optimum is interior. On the other hand, under weak learnability, eq. (6.4) provided that the dual optimum has zeros at every coordinate. As will be made clear in section 8, the primal and dual weights have the following dichotomy: either the margin $\langle a_i, \lambda\rangle$ goes to infinity and $(\psi_A)_i$ goes to zero, or the margin stays finite and $(\psi_A)_i$ goes to some positive value.

Theorem 7.6. *Suppose $A \neq \mathbf{0}_{m\times n}$, $g \in \mathbb{G}$, and the infimum of eq. (2.1) is attainable. Then there exists a (compact) tightest axis-aligned retangle $\mathcal{C}$ containing the initial level set, and f is strongly convex with modulus $c > 0$ over $\mathcal{C}$. Finally, $\gamma(A, -\nabla f(\mathcal{C})) > 0$, and for all t,*

$$f(A\lambda_t) - \bar{f}_A \leq (f(\mathbf{0}_m) - \bar{f}_A)\left(1 - \frac{c\gamma(A, -\nabla f(\mathcal{C}))^2}{\eta f(A\lambda_0)}\right)^t.$$

In other words, $t \leq \frac{\eta f(A\lambda_0)}{c\gamma(A, -\nabla f(\mathcal{C}))^2}\ln(\frac{1}{\epsilon})$ iterations suffice to reach error ϵ. The appearance of a modulus of strong convexity c (i.e., a lower bound on the eigenvalues of the Hessian of f) may seem surprising, and sketching the proof illuminates its appearance and subsequent function.

When the infimum is attainable, every margin $\langle a_i, \lambda \rangle$ converges to some finite value. In fact, they all remain bounded: (7.2) provides that no halfspace contains all points, so if one margin becomes positive and large, another becomes negative and large, giving a terrible objective value. But objective values never increase with coordinate descent. To finish the proof, strong convexity (i.e., quadratic lower bounds in the primal) grants quadratic upper bounds in the dual, which can be used to bound the dual distance in proposition 5.2, and yield the desired convergence rate. This approach fails under weak learnability—some primal weights grow unboundedly, all dual weights shrink to zero, and no compact set contains all margins.

8   General Convergence Rate

The final characterization encodes two principles: the rows of A may be partitioned into two matrices A_0, A_+ which respectively satisfy theorem 6.1 and theorem 7.1, and that these two subproblems affect the optimization problem essentially independently.

Theorem 8.1. *Let $A_0 \in \mathbb{R}^{z \times n}$, $A_+ \in \mathbb{R}^{p \times n}$, and $g \in \mathbb{G}$ be given. Set $m := z + p$, and $A \in \mathbb{R}^{m \times n}$ to be the matrix obtained by stacking A_0 on top of A_+. The following conditions are equivalent:*

$$(\exists \lambda \in \mathbb{R}^n \bullet A_0 \lambda \in \mathbb{R}^z_{++} \wedge A_+ \lambda = \mathbf{0}_p) \wedge (\forall \lambda \in \mathbb{R}^n \bullet A_+ \lambda \notin \mathbb{R}^p_+ \setminus \{\mathbf{0}_p\}), \tag{8.2}$$

$$(\inf_{\lambda \in \mathbb{R}^n} f(A\lambda) = \inf_{\lambda \in \mathbb{R}^n} f(A_+\lambda)) \wedge (\inf_{\lambda \in \mathbb{R}^n} f(A_0\lambda) = 0) \wedge f \circ A_+ \text{ has minimizers}, \tag{8.3}$$

$$\psi_A = \begin{bmatrix} \psi_{A_0} \\ \psi_{A_+} \end{bmatrix} \text{ with } \psi_{A_0} = \mathbf{0}_z \wedge \psi_{A_+} \in \mathbb{R}^p_{++}, \tag{8.4}$$

$$(\Phi_{A_0} = \{\mathbf{0}_z\}) \wedge (\Phi_{A_+} \cap \mathbb{R}^p_{++} \neq \emptyset) \wedge (\Phi_A = \Phi_{A_0} \times \Phi_{A_+}). \tag{8.5}$$

To see that any matrix A falls into one of the three scenarios here, fix a loss function g, and recall from theorem 3.1 that ψ_A is unique. In particular, the set of zero entries in ψ_A exactly specifies which of the three scenarios hold, the current scenario allowing for simultaneous positive and zero entries. Although this reasoning made use of ψ_A, note that it is A which dictates the behavior: in fact, as is shown in remark I.1 of the full version, the decomposition is unique.

Returning to theorem 8.1, the geometry of fig. 2c is provided by (8.2) and (8.5). The analysis will start from (8.3), which allows the primal problem to be split into two pieces, which are then individually handled precisely as in the preceding sections. To finish, (8.5) will allow these pieces to be stitched together.

Theorem 8.6. *Suppose $A \neq \mathbf{0}_{m \times n}$, $g \in \mathbb{G}$, $\psi_A \in \mathbb{R}^m_+ \setminus \mathbb{R}^m_{++} \setminus \{\mathbf{0}_m\}$, and the notation from theorem 8.1. Set $w := \sup_t \|\nabla f(A_+\lambda_t) + \mathrm{P}^1_{\Phi_{A_+}}(-\nabla f(A_+\lambda_t))\|_1$. Then $w < \infty$, and there exists a tightest cube $\mathcal{C}_+$ so that $\mathcal{C}_+ \supseteq \{x \in \mathbb{R}^p : f(x) \leq f(A\lambda_0)\}$, and let $c > 0$ be the modulus of strong convexity of f over $\mathcal{C}_+$. Then $\gamma(A, \mathbb{R}^z_+ \times -\nabla f(\mathcal{C}_+)) > 0$, and for all t, $f(A\lambda_t) - \bar{f}_A \leq 2f(A\lambda_0)/\left((t+1)\min\left\{1, \gamma(A, \mathbb{R}^z_+ \times -\nabla f(\mathcal{C}_+))^2/((\beta + w/(2c))^2\eta)\right\}\right)$.*

(In the case of the logistic loss, $w \leq \sup_{x \in \mathbb{R}^m} \|\nabla f(x)\|_1 \leq m$.)

As discussed previously, the bounds deteriorate to $\mathcal{O}(1/\epsilon)$ because the finite and infinite margins sought by the two pieces A_0, A_+ are in conflict. For a beautifully simple, concrete case of this, consider the following matrix, due to Schapire [11]:

$$S := \begin{bmatrix} -1 & +1 \\ +1 & -1 \\ +1 & +1 \end{bmatrix}.$$

The optimal solution here is to push both coordinates of λ unboundedly positive, with margins approaching $(0, 0, \infty)$. But pushing any coordinate λ_i too quickly will increase the objective value, rather than decreasing it. In fact, this instance will provide a lower bound, and the mechanism of the proof shows that the primal weights grow extremely slowly, as $\mathcal{O}(\ln(t))$.

Theorem 8.7. *Using the logistic loss and exact line search, for any $t \geq 1$, $f(S\lambda_t) - \bar{f}_S \geq 1/(8t)$.*

References

[1] Robert E. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, July 1990.

[2] Yoav Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2):256–285, 1995.

[3] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

[4] Leo Breiman. Prediction games and arcing algorithms. *Neural Computation*, 11:1493–1517, October 1999.

[5] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28, 1998.

[6] Peter J. Bickel, Yaacov Ritov, and Alon Zakai. Some theory for generalized boosting algorithms. *Journal of Machine Learning Research*, 7:705–732, 2006.

[7] Z. Q. Luo and P. Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72:7–35, 1992.

[8] Michael Collins, Robert E. Schapire, and Yoram Singer. Logistic regression, AdaBoost and Bregman distances. *Machine Learning*, 48(1-3):253–285, 2002.

[9] Gunnar Rätsch, Sebastian Mika, and Manfred K. Warmuth. On the convergence of leveraging. In *NIPS*, pages 487–494, 2001.

[10] Indraneel Mukherjee, Cynthia Rudin, and Robert Schapire. The convergence rate of AdaBoost. In *COLT*, 2011.

[11] Robert E. Schapire. The convergence rate of AdaBoost. In *COLT*, 2010.

[12] Robert E. Schapire, Yoav Freund, Peter Barlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *ICML*, pages 322–330, 1997.

[13] Gunnar Rätsch and Manfred K. Warmuth. Maximizing the margin with boosting. In *COLT*, pages 334–350, 2002.

[14] Manfred K. Warmuth, Karen A. Glocer, and Gunnar Rätsch. Boosting algorithms for maximizing the soft margin. In *NIPS*, 2007.

[15] Shai Shalev-Shwartz and Yoram Singer. On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms. In *COLT*, pages 311–322, 2008.

[16] Llew Mason, Jonathan Baxter, Peter L. Bartlett, and Marcus R. Frean. Functional gradient techniques for combining hypotheses. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 221–246, Cambridge, MA, 2000. MIT Press.

[17] Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[18] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of Convex Analysis*. Springer Publishing Company, Incorporated, 2001.

[19] Jonathan Borwein and Adrian Lewis. *Convex Analysis and Nonlinear Optimization*. Springer Publishing Company, Incorporated, 2000.

[20] Robert E. Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.

[21] George B. Dantzig and Mukund N. Thapa. *Linear Programming 2: Theory and Extensions*. Springer, 2003.

[22] Adi Ben-Israel. Motzkin's transposition theorem, and the related theorems of Farkas, Gordan and Stiemke. In M. Hazewinkel, editor, *Encyclopaedia of Mathematics, Supplement III*. 2002.

A Notation

To aid in the reading of the proofs, this section summarizes notation used throughout the manuscript.

Symbol	Comment
$\mathbb{R}^m$	m-dimensional vector space over the reals.
$\mathbb{R}^m_+$	Non-negative m-dimensional real vectors.
$\mathrm{int}(S)$	The interior of set S.
$\mathbb{R}^m_{++}$	Positive m-dimensional real vectors, i.e. $\mathrm{int}(\mathbb{R}^m_+)$.
$\mathbf{0}_m$	m-dimensional Vector of all zeros.
$\mathbf{e}_i$	Indicator vector: 1 at coordinate i, 0 elsewhere. Context will provide the ambient dimension.
$\mathrm{Im}(A)$	Image of linear operator A.
$\mathrm{Ker}(A)$	Kernel of linear operator A.
I_n	Identity matrix with rank n.

ι_S	Indicator function on a set S:

$$\iota_S(x) := \begin{cases} 0 & x \in S, \\ \infty & x \notin S. \end{cases}$$

$\mathrm{dom}(h)$	Domain of convex function h, i.e. the set $\{x \in \mathbb{R}^m : h(x) < \infty\}$.
h^*	The Fenchel conjugate of h:

$$h^*(\phi) = \sup_{x \in \mathrm{dom}(h)} \langle \phi, x \rangle - h(x).$$

For any h considered in this document, h^* is closed and convex [18, Theorem E.1.1.2]. A key property is that, when both gradients exist, $\nabla h^*(\nabla h(x)) = x$ (see for instance [18, Corollary E.1.4.4] for the generalization to subgradients): the Fenchel conjugate thus provides an inverse gradient map. For a beautiful description of Fenchel conjugacy, please see [18, Section E.1.2].

0-coercive	A closed convex function f with all level sets compact is called 0-coercive; for strictly convex functions, this is equivalent to the infimum being attainable (cf. proposition H.1). For a thorough treatment of 0-coercivity, please see Proposition B.3.2.4 and Definition B.3.2.5 of [18].

Symbol	Comment
$\mathbb{G}_0$	Basic loss class under consideration (cf. section 2).
$\mathbb{G}$	Refined loss class for which convergence rates are established (cf. section 5).
η, β	Parameters corresponding to some $g \in \mathbb{G}$ (cf. section 5).
Φ_A	The general dual feasibility set: $\Phi_A := \mathrm{Ker}(A^\top) \cap \mathbb{R}^m_+$.
$\gamma(A, S)$	Generalization of classical weak learning rate (cf. section 4).
$\bar{f}_A$	The minimal objective value of $f \circ A$: $\bar{f}_A := \inf_\lambda f(A\lambda)$. In general, this value is not attainable.
ψ_A	Dual optimum. This value always exists, is unique, and non-negative (cf. section 3).
P^p_S	l^p projection onto closed nonempty convex set S. When $p \in \{1, \infty\}$, this is not satisfied uniquely, thus choose any element (cf. section 4).
D^p_S	l^p distance to closed nonempty convex set S: $\mathsf{D}^p_S(\phi) := \|\phi - \mathsf{P}^p_S(\phi)\|_p$.

B A Few Key Results from the Literature

The first three results are canonical theorems of alternatives. Incidentally, Gordan's theorem, to the knowledge of the author, is the oldest theorem of alternatives [see 21, Bibliographic notes, Section 5 of Chapter 2]. A streamlined presentation, using a related optimization problem (which can nearly be written as $f \circ A$ from this manuscript), can be found in [19, Theorem 2.2.6].

Theorem B.1 (Gordan, cf. Theorem 2.2.1 of [19])**.** *For any $A \in \mathbb{R}^{m \times n}$, exactly one of the following situations holds:*

$$\exists \phi \in \mathbb{R}^m_+ \setminus \{\mathbf{0}_m\} \centerdot A^\top \phi = \mathbf{0}_n; \tag{B.2}$$

$$\exists \lambda \in \mathbb{R}^n \centerdot A\lambda \in \mathbb{R}^m_{++}. \tag{B.3}$$

A geometric interpretation is as follows. Take the rows of A to be m points in $\mathbb{R}^n$. Then there are two possibilities: either there exists an open homogeneous halfspace containing all points, or their convex hull contains the origin.

Next is Stiemke's Theorem of Alternatives.

Theorem B.4 (Stiemke, cf. Exercise 2.2.8 of [19]). *For any $A \in \mathbb{R}^{m \times n}$, exactly one of the following situations holds:*

$$\exists \phi \in \mathbb{R}_{++}^m \,\textbf{.}\, A^\top \phi = \mathbf{0}_n; \tag{B.5}$$

$$\exists \lambda \in \mathbb{R}^n \,\textbf{.}\, A\lambda \in \mathbb{R}_+^m \setminus \{\mathbf{0}_m\}. \tag{B.6}$$

The geometric interpretation here is that either there exists a closed homogeneous halfspace containing all m points, with at least one point interior to the halfspace, or the relative interior of the convex hull of the points contains the origin (for the connection to relative interiors, see for instance [18, Remark A.2.1.4]).

Third, a version of Motzkin's Transposition Theorem, which can encode the theorems of alternatives due to Farkas, Stiemke, and Gordan [22].

Theorem B.7 (Motzkin, cf. Theorem 2.16 of [21]). *For any $B \in \mathbb{R}^{z \times n}$ and $C \in \mathbb{R}^{p \times n}$, exactly one of the following situations holds:*

$$\exists \lambda \in \mathbb{R}^n \,\textbf{.}\, B\lambda \in \mathbb{R}_{++}^z \wedge C\lambda \in \mathbb{R}_+^p, \tag{B.8}$$

$$\exists \phi_B \in \mathbb{R}_+^z \setminus \{\mathbf{0}_z\}, \phi_C \in \mathbb{R}_+^p \,\textbf{.}\, B^\top \phi_B + C^\top \phi_C = \mathbf{0}_n. \tag{B.9}$$

For this geometric interpretation, suppose a matrix $A \in \mathbb{R}^{m \times n}$, broken into two submatrices $B \in \mathbb{R}^{z \times n}$ and $C \in \mathbb{R}^{p \times n}$, with $z + p = m$; again, consider the rows of A as m points in $\mathbb{R}^n$. The first possibility is that there exists a closed homogeneous halfspace containing all m points, the z points corresponding to B being interior to the halfspace. Otherwise, the origin can be written as a convex combination of these m points, with positive weight on at least one element of B.

The next result is a form of the Fenchel-Young inequality, simplified for differentiability (the "if" can be strengthened to "iff" via subgradients).

Proposition B.10 (Fenchel-Young Inequality, Proposition 3.3.4 in [19]). *For any convex function f and $x \in \mathrm{dom}(f)$, $\phi \in \mathrm{dom}(f^*)$,*

$$f(x) + f^*(\phi) \geq \langle x, \phi \rangle,$$

with equality if $\phi = \nabla f(x)$.

Next, a lemma to convert single-step convergence results into general convergence results. Results of this sort are widely used in convex analysis.

Lemma B.11 (Lemma 20 from [15]). *Suppose $\infty > a_0 \geq a_1 \geq \ldots \geq 0$ and $a_{i+1} \leq a_i - r a_i^2$ where $r \in (0, (2a_0)^{-1}]$. Then $a_i \leq \frac{1}{r(i+2)}$.*

Proof. To reduce this to the exact setting of Lemma 20 from [15], set $\epsilon_i = a_{i-1}/a_0$, so now $1 = \epsilon_1 \geq \epsilon_2 \geq \ldots$ with corresponding inequality

$$\epsilon_{i+1} \leq \epsilon_i - a_0 r \epsilon_i^2.$$

Since $a_0 r \in (0, 1/2]$, the lemma may be invoked, yielding $\epsilon_i \leq (a_0 r (i+1))^{-1}$, which can be rearranged to give the result. $\square$

Although strong convexity in the primal grants the existence of a lower bounding quadratic, it grants upper bounds in the dual. The following result is also standard in convex analysis, see for instance the proof of [18, Theorem E.4.2.2].

Lemma B.12 (Lemma 18 from [15]). *Let h be strongly convex over compact convex set S with modulus c. Then for any $\phi_1, \phi_1 + \phi_2 \in \nabla f(S)$,*

$$f^*(\phi_1 + \phi_2) - f^*(\phi_1) \leq \langle \nabla f^*(\phi_1), \phi_2 \rangle + \frac{1}{2c} \|\phi_2\|_2^2.$$

C   Basic Properties of $g \in \mathbb{G}_0$

Lemma C.1. *Let any $g \in \mathbb{G}_0$ be given. Then g is strictly convex, $g > 0$, g strictly decreases ($g' < 0$), and g' strictly increases. Lastly, $\lim_{x \to -\infty} g(x) = \infty$.*

Proof. (Strict convexity and g' strictly increases.) For any $x < y$,

$$g'(y) = g'(x) + \int_x^y g''(t)dt \geq g'(x) + (y - x)\inf_{t \in [x,y]} g''(t) > g'(x),$$

and thus g' is strictly monotone, granting strict convexity [see 18, Theorem B.4.1.4].

(g strictly decreases, i.e. $g' < 0$.) Suppose there exists $x < y$ with $g(y) > g(x)$. By convexity,

$$g(x) \geq g(y) + g'(y)(x - y) = g(y) - g'(y)(y - x)$$

$$\implies \qquad g'(y) \geq \frac{g(y) - g(x)}{y - x} =: c > 0.$$

Thus, for any $z > y$,

$$g(z) \geq g(y) + g'(y)(z - y) \geq g(y) + c(z - y),$$

which contradicts $\lim_{x \to \infty} g(x) = 0$. Thus g is non-increasing. Furthermore, it is decreasing, since if there existed $x < y$ with $g(x) = g(y)$, strict convexity would grant $g((x + y)/2) < g(y)$, which contradicts the non-increasing property.

($g > 0$.) If there existed y with $g(y) \leq 0$, then the strict decreasing property would invalidate $\lim_{x \to \infty} g(x) = 0$.

($\lim_{x \to -\infty} g(x) = \infty$.) Let any sequence $\{c_i\}_1^\infty \downarrow -\infty$ be given; the result follows by convexity and $g' < 0$, since

$$\lim_{i \to \infty} g(c_i) \geq \lim_{i \to \infty} g(c_1) + g'(c_1)(c_i - c_1) = \infty. \qquad \square$$

Lemma C.2. *Let $g \in \mathbb{G}_0$ be given. Then g^* is continuously differentiable on $\mathrm{int}(\mathrm{dom}(g^*))$, strictly convex, and either $\mathrm{dom}(g^*) = (-\infty, 0]$ or $\mathrm{dom}(g^*) = [b, 0]$ where $b < 0$. Furthermore, g^* has the following form:*

$$g^*(\phi) \in \begin{cases} (-g(0), \infty] & \phi < g'(0), \\ -g(0) & \phi = g'(0), \\ (-g(0), 0) & \phi \in (g'(0), 0), \\ 0 & \phi = 0, \\ \infty & \phi > 0. \end{cases}$$

Proof. g^* is strictly convex because g is differentiable, and g^* is continuously differentiable on $\mathrm{int}(\mathrm{dom}(g^*))$ because g is strictly convex [see 18, Theorems E.4.1.1, E.4.1.2].

Next, when $\phi > 0$: $\lim_{x \to \infty} g(x) = 0$ grants the existence of y such that for any $x \geq y$, $g(x) \leq 1$, thus

$$g^*(\phi) = \sup_x \phi x - g(x) \geq \sup_{x \geq y} \phi x - 1 = \infty.$$

(Since $g > 0$, this precludes the possibility of $\infty - \infty$)

Take $\phi = 0$; then

$$g^*(\phi) = \sup_x -g(x) = -\inf_x g(x) = 0.$$

When $\phi = g'(0)$, by the Fenchel-Young inequality (proposition B.10)

$$g^*(\phi) = g^*(g'(0)) = 0 \cdot g'(0) - g(0) = -g(0).$$

Moreover, by [18, Corollary E.1.4.3] $\nabla g^*(g'(0)) = 0$, which combined with strict convexity of g^* means $g'(0)$ minimizes g^*. g^* is closed [18, Theorem E.1.1.2], which combined with the above gives that $\mathrm{dom}(g^*) = (-\infty, 0]$ or $\mathrm{dom}(g^*) = [-b, 0]$ for some $b < 0$, and the rest of the form of g^*. $\qquad \square$

12

Lemma C.3. *Let any $g \in \mathbb{G}_0$ be given. Then the corresponding f is strictly convex, twice differentiable, and $f' < \mathbf{0}_m$. Furthermore, $\mathrm{dom}(f^*) = \mathrm{dom}(g^*)^m \subseteq \mathbb{R}^m_-$, $f^*(\mathbf{0}_m) = 0$, f^* is strictly convex, f^* is continuously differentiable on the interior of its domain, and finally $f^*(\phi) = \sum_{i=1}^m g^*(\phi_i)$.*

Proof. First,

$$f^*(\phi) = \sup_{x \in \mathbb{R}^m} \langle \phi, x \rangle - f(x) = \sup_{x \in \mathbb{R}^m} \sum_{i=1}^m x_i \phi_i - g(x_i) = \sum_{i=1}^m g^*(\phi_i).$$

The remaining properties follow from properties of g and g^* (cf. lemma C.1 and lemma C.2). $\qquad\square$

D   Deferred Material from Section 3

Proof of theorem 3.1. Writing the objective as two Fenchel problems,

$$p := \inf_\lambda f(A\lambda) + \iota_{\mathbb{R}^n}(\lambda),$$

$$d := \sup_\phi -f^*(-\phi) - \iota^*_{\mathbb{R}^n}(A^\top \phi).$$

Since $\mathrm{cont}(f) = \mathbb{R}^m$ (set of points where f is continuous) and $\mathrm{dom}(\iota_{\mathbb{R}^n}) = \mathbb{R}^n$, it follows that $A\mathrm{dom}(\iota_{\mathbb{R}^n}) \cap \mathrm{cont}(f) = \mathrm{Im}(A) \neq \emptyset$, thus by [19, Theorem 3.3.5], $p = d$. Moreover, since $p \leq f(\mathbf{0}_m)$ and $d \geq -f^*(\mathbf{0}_m) = 0$, the optimum is finite, and thus the same theorem grants that it is attainable in the dual.

To finish, note for any $\lambda \in \mathbb{R}^n$ that

$$\iota^*_{\mathbb{R}^n}(\lambda) = \sup_{\mu \in \mathbb{R}^n} \langle \lambda, \mu \rangle - \iota_{\mathbb{R}^n}(\mu) = \iota_{\{\mathbf{0}_n\}}(\lambda).$$

Thus $\phi \in \mathrm{Ker}(A^\top)$; and by lemma C.3, $\mathrm{dom}(f^*) \subseteq \mathbb{R}^m_-$, and thus $\phi \in \Phi_A = \mathrm{Ker}(A^\top) \cap \mathbb{R}^m_+$.

The fact that $f^*(\phi) = \sum_i g^*((\phi)_i)$ was proved in lemma C.3.

The uniqueness of ψ_A was established by Collins et al. [8, Theorem 1], however a direct argument is as follows by the strict convexity of f^* (cf. lemma C.3). Specifically, this means the optimum ψ is unique, for if there were some other optimal $\psi' \neq \psi$, the point $(\psi + \psi')/2$ is dual feasible and has strictly smaller objective value, a contradiction. $\qquad\square$

E   Deferred Material from Section 4

The proof of proposition 4.3 is technical, but the central strategy is straightforward. First note that $\gamma(A, S)$ can be rewritten as

$$\gamma(A, S) = \inf_{S \backslash \mathrm{Ker}(A^\top)} \frac{\|A^\top \phi\|_\infty}{\|\phi - \mathsf{P}^1_{S \cap \mathrm{Ker}(A^\top)}(\phi)\|_1}$$

$$= \inf_{S \backslash \mathrm{Ker}(A^\top)} \frac{\|A^\top(\phi - \mathsf{P}^1_{S \cap \mathrm{Ker}(A^\top)}(\phi))\|_\infty}{\|\phi - \mathsf{P}^1_{S \cap \mathrm{Ker}(A^\top)}(\phi)\|_1}$$

$$= \inf \left\{ \|A^\top v\|_\infty : \forall v \in \mathbb{R}^m \centerdot \exists \phi \in S \centerdot v = \phi - \mathsf{P}^1_{S \cap \mathrm{Ker}(A^\top)}(\phi), \|v\|_1 = 1 \right\},$$

where the second equivalence used $A^\top \mathsf{P}^1_{S \cap \mathrm{Ker}(A^\top)}(\phi) = \mathbf{0}_n$.

In the final form, $v \notin \mathrm{Ker}(A^\top)$, and so $A^\top v \neq \mathbf{0}_n$; that is to say, the infimand is positive for every element of its domain. The difficulty is that the domain of the infimum, written in this way, is not obviously closed; thus one can not simply assert the infimum is attainable and positive.

The goal then will be to reparameterize the infimum as having a compact domain. For technical convenience, the result will be mainly proved for the l^2 norm (where projections are very well-behaved), and norm equivalence will provide the final result.

Lemma E.1. *Given $A \neq \mathbf{0}_{m \times n}$ and a polyhedron S with nonempty interior and $S \cap \mathrm{Ker}(A^\top) \neq \emptyset$,*

$$\inf \left\{ \frac{\|A^\top(\phi - \mathsf{P}^2_{S \cap \mathrm{Ker}(A^\top)}(\phi))\|_2}{\|\phi - \mathsf{P}^2_{S \cap \mathrm{Ker}(A^\top)}(\phi)\|_2} : \phi \in S \setminus \mathrm{Ker}(A^\top) \right\} \tag{E.2}$$

Proof. First dispense with a technical consideration: since $A \neq \mathbf{0}_{m \times n}$, then $\ker(A^\top) \neq \mathbb{R}^m$, meaning $S \setminus \ker(A^\top) \neq \emptyset$ (since S has interior points). In particular, the domain of the infimum is nonempty. Furthermore, since $S \cap \mathrm{Ker}(A^\top)$ is nonempty, projections onto it are well defined.

Let $(\phi_i^{(1)})_{i=1}^\infty$ be a minimizing sequence for (E.2), and for convenience set $K := \mathrm{Ker}(A^\top)$.

Amongst the family of all polyhedral subsets of $S \cap K$, notice that some receive an infinite subsequence of the projected elements $(\mathsf{P}^2_{S \cap K}(\phi_i^{(1)}))_{i=1}^\infty$. Partition this family by affine dimension; since the polyhedron $S \cap K$ is in this family, there is a subpolyhedron with affine dimension $\dim(S \cap K)$ receiving infinitely many projections. Now designate E to be any polyhedral subset of smallest dimension amongst those receiving an infinite subsequence of projections, and label E's subsequence $(\phi_i^{(2)})_{i=1}^\infty$.

The strict subfaces of E (i.e., the faces of E which are not E itself) are also polyhedral subsets of $S \cap K$. Their dimension, however, is strictly smaller than E's dimension, meaning each subface receives only finitely many projection elements. Since there are only finitely many subfaces, produce a subsequence $(\phi_i^{(3)})_{i=1}^\infty$ by deleting all projections onto strict subfaces. Crucially, as precisely the projections onto the relative boundary of E were cut, it now holds that $\mathsf{P}^2_{S \cap K}(\phi_i^{(3)}) \in \mathrm{ri}(E)$, the relative interior of E. Lastly, note that this is still a minimizing sequence for (E.2).

Next note that the normal cone $N_E(x)$ to E at a relatively interior point $x \in \mathrm{ri}(E)$ is a subspace; in particular, it is

$$(\mathrm{aff}(E) - \{x\})^\perp,$$

the orthogonal complement of the subspace corresponding to the affine hull of E. (This can be verified by considering any point whose projection is relatively interior to E; if it were not orthogonal to $\mathrm{aff}(E)$, then one could choose a point in a relative neighboorhood of x, but closer to the projection onto $\mathrm{aff}(E)$, thus contradicting the choice of projection element as the closest element in E to x. For related results, please see Hiriart-Urruty and Lemaréchal [18, sections A.5.2, A.5.3]).

Since S is a polyhedron, it has a representation

$$S = \{x \in \mathbb{R}^m : \forall i \centerdot g_i(x) \leq 0\},$$

where $\{g_i\}_1^k$ is a finite collection of affine functions. For any point $x \in E$, define the *active set* $\mathcal{I}_x$ to be the constraints of S which are exactly met for x:

$$\mathcal{I}_x := \{i : g_i(x) = 0\}.$$

If x, y are in $\mathrm{ri}(E)$, then $\mathcal{I}_x = \mathcal{I}_y$. To see this, assume contradictorily that $\mathcal{I}_x \neq \mathcal{I}_y$; without loss of generality, let g_k be some affine manifold with $g_k(x) < 0$ but $g_k(y) = 0$. By definition of relative interior, there exists $\epsilon > 0$ so that $B(y, \epsilon) \cap \mathrm{aff}(E) \subseteq E$. Thanks to this, the point $y' := y + 2^{-1}\epsilon(y - x)/\|y - x\|_2$, which is along the line through x and y, is still in E. On the other hand, the hyperplane corresponding to g_k must separate x from y', and in particular $g_k(y') > 0$, meaning $y' \notin S$, contradicting the fact that $E \subseteq S$. Thus every element of $\mathrm{ri}(E)$ has the same active set.

For any $x \in E$, the distance from x to any hyperplane defining S, but not in $\mathcal{I}_x$, is positive. Since S is defined by finitely many hyperplanes, there exists $\delta > 0$ such that the ball $B(x, \delta)$ intersects only those hyperplanes in $\mathcal{I}_x$. Thus define the set

$$\mathcal{C}_x := \{v \in \mathbb{R}^m : \|v\|_2 \leq \delta, \forall j \in \mathcal{I}_x \centerdot g_j(x + v) \leq 0, v \in N_E(x)\},$$

which is the collection of all projection directions to x from points within $S \cap B(x, \delta)$.

Since $N_E(x)$ and $\mathcal{I}_x$ are the same for all x in the relative interior of $\mathcal{C}_x$, all that may differ across points is the choice of δ; accordingly, for any x, y in $\mathrm{ri}(E)$, $\mathcal{C}_x = a\mathcal{C}_y$ for some $a > 0$.

14

Finally, observe that projections onto x from farther than the constructed δ away must simply be rescalings of the elements of $\mathcal{C}_x$; where this not the case, S would be nonconvex. Symbolically,

$$\mathcal{C}_x \subset \{\phi - x : \phi \in S, \mathsf{P}^2_{K \cap S}(\phi) = x\} \subseteq \mathbb{R}_+ \mathcal{C}_x;$$

Notice that

$$\mathbb{R}_+ \mathcal{C}_x = \{v \in \mathbb{R}^m : \forall j \in \mathcal{I}_x \cdot g_j(x + v) \le 0, v \in N_E(x)\};$$

that is to say, $\mathbb{R}_+ \mathcal{C}_x$ is a closed polyhedral cone.

Using all these facts,

$$
\inf\left\{\frac{\|A^\top \phi\|_2}{\|\phi - \mathsf{P}^2_{S \cap K}(\phi)\|_2} : \phi \in S \setminus K\right\} = \inf\left\{\frac{\|A^\top(\phi_i^{(3)} - \mathsf{P}^2_E(\phi_i^{(3)}))\|_2}{\|\phi_i^{(3)} - \mathsf{P}^2_E(\phi_i^{(3)})\|_2} : i \in \mathbb{Z}_+\right\}
$$

$$
\ge \inf\left\{\frac{\|A^\top v\|_2}{\|v\|_2} : x \in \mathrm{ri}(E), v \in \mathcal{C}_x \setminus \{\mathbf{0}_m\}\right\}
$$

$$
\ge \inf\left\{\frac{\|A^\top \phi\|_2}{\|\phi - \mathsf{P}^2_{S \cap K}(\phi)\|_2} : \phi \in S \setminus K\right\},
$$

where the presence of the final inequality provides that this is in fact a chain of equalities. Fixing some $x_0 \in \mathrm{ri}(E)$ and noting that the infimand ignores the length of v,

$$
\inf\left\{\frac{\|A^\top v\|_2}{\|v\|_2} : x \in \mathrm{ri}(E), v \in \mathcal{C}_x \setminus \{\mathbf{0}_m\}\right\} = \inf\left\{\frac{\|A^\top v\|_2}{\|v\|_2} : x \in \mathrm{ri}(E), v \in \mathbb{R}_+ \mathcal{C}_x \setminus \{\mathbf{0}_m\}\right\}
$$

$$
= \inf\left\{\frac{\|A^\top v\|_2}{\|v\|_2} : v \in \mathbb{R}_+ \mathcal{C}_{x_0} \setminus \{\mathbf{0}_m\}\right\}
$$

$$
= \inf\left\{\|A^\top v\|_2 : v \in \mathbb{R}_+ \mathcal{C}_{x_0} \cap B(1,0)\right\}.
$$

Since it was provided above that $\mathbb{R}_+ \mathcal{C}_{x_0}$ is a closed polyhedral cone, the domain of this final infimum is a compact set. As the infimand is continuous, this infimum attains a minimizer $\bar{v}$, and there must exist $c > 0$ so that $x_0 + c\bar{v} \in S$ and $\mathsf{P}^2_E(x_0 + c\bar{v}) = x_0 \in E \subseteq K$. But then $c\bar{v} \notin \mathrm{Ker}(A^\top)$, meaning $\|A^\top \bar{v}\|_2 > 0$, and the infimum is positive. $\square$

This proof has an interesting byproduct; to compute this quantity, one can simply focus on orthogonal projections to some very bad subpolyhedron's relative interior.

Proof of proposition 4.3. For the upper bound, note as in the proof of lemma E.1 that $S \cap \mathrm{Ker}(A^\top)$ is nonempty and the infimand is positive for every element of the domain, so the infimum is finite. For the lower bound, by lemma E.1 and norm equivalence,

$$
\gamma_p(A) = \inf_{\phi \in S \setminus \mathrm{Ker}(A^\top)} \frac{\|A^\top \phi\|_\infty}{\inf_{\psi \in S \cap \mathrm{Ker}(A^\top)} \|\phi - \psi\|_1}
$$

$$
\ge \left(\frac{1}{\sqrt{mn}}\right) \inf_{\phi \in S \setminus \mathrm{Ker}(A^\top)} \frac{\|A^\top \phi\|_2}{\inf_{\psi \in S \cap \mathrm{Ker}(A^\top)} \|\phi - \psi\|_2} > 0. \qquad \square
$$

F Deferred Material from Section 5

Remark F.1. This remark develops bounds on the quantities η, β for the logistic loss $g = \ln(1 + \exp(-\cdot))$. First note that the initial level set $S_0 := \{x \in \mathbb{R}^m : f(X) \le f(A\lambda_0)\}$ is contained within a cube $[b, \infty)^m$, where $b \ge -m\ln(2)$; this follows since $f(A\lambda_0) = f(\mathbf{0}_m) = m\ln(2)$, whereas $g(-m\ln(2)) = \ln(1 + \exp(-(-m\ln(2)))) \ge m\ln(2)$.

The analysis will be written with respect to b. Let any $x \in [b, \infty)$ be given, and note $g' = -(1 + \exp(\cdot))^{-1}$, and $g'' = \exp(\cdot)(1 + \exp(\cdot))^{-2}$.

To start determining η, note $1 \le 1 + \exp(-x) \le 1 + \exp(-b)$. Set $c_1 := \ln(1 + \exp(-b))$; since $\ln$ is concave, it follows that, for all $z \in [1, 1 + \exp(-b)]$, the secant line through $(1, 0)$ and $(1 + \exp(-b), c_1)$ is a lower bound:

$$
\ln(z) \ge \left(\frac{c_1 - 0}{1 + \exp(-b) - 1}\right) z - \frac{c_1 - 0}{1 + \exp(-b) - 1} = c_1 e^b(z - 1).
$$

Thus, setting $\eta := 1/(c_1 \exp(b))$, for $x \in [b, \infty)$, $\ln(1 + \exp(-x)) \geq \exp(-x)/\eta$. Furthermore

$$\frac{g''(x)}{g(x)} = \frac{e^{-x}}{(1 + e^{-x})^2 \ln(1 + e^{-x})} \leq \frac{\eta e^{-x}}{e^{-x}},$$

and thus $g''(x) \leq \eta g(x)$.

And for $g(x) \leq -\beta g'(x)$, using $\ln(x) \leq x - 1$,

$$\frac{-g'(x)}{g(x)} = \frac{\frac{\exp(-x)}{1+\exp(-x)}}{\ln(1 + \exp(-x))} \geq \frac{\frac{\exp(-x)}{1+\exp(-x)}}{\exp(-x)} \geq \frac{1}{1 + \exp(-b)}.$$

That is, it suffices to set $\beta := 1 + \exp(-b)$. $\diamond$

The next result, which is standard in the optimization literature (the presentation here is resembles one due to Boyd and Vandenberghe [17, section 9.4]), provides a lower bound on the improvement due to a single single descent iteration. In particular, one can simply choose a step size based on the minimizer of an upper bounding quadratic. As discussed in section 2, instead using a standard iterative line search will only degrade the presented bounds by a constant. Note however that even though the quantities in the following lemma are known to the algorithm at runtime, it is preferable in practice to use an iterative line search; this result can thus be taken as providing the existence of good line search choices, regardless of existence and computational considerations of an exact line search minimizer.

Lemma F.2. *Fix any t and $g \in \mathbb{G}_0$ (with corresponding f). Suppose the line search chooses $\alpha_{t+1} := -\langle A^\top \nabla f(A\lambda_t), \mathbf{e}_{j_{t+1}} \rangle / (\eta f(A\lambda_t))$. Then*

$$f(A\lambda_t) - f(A\lambda_{t+1}) \geq \frac{\|A^\top \nabla f(A\lambda_t)\|_\infty^2}{2\eta f(A\lambda_t)}.$$

Proof. Define a level set for the line search:

$$L_{t+1} := \{\alpha \in \mathbb{R} : (f \circ A)(\lambda_t + \alpha \mathbf{e}_{j_{t+1}}) \leq (f \circ A)(\lambda_t)\}.$$

Since $\alpha \mapsto (f \circ A)(\lambda_t + \alpha \mathbf{e}_{j_{t+1}})$ is convex, L_{t+1} is convex. By Taylor's theorem, for any $\alpha \in L_{t+1}$,

$$(f \circ A)(\lambda_t + \alpha \mathbf{e}_{j_{t+1}}) \leq f(A\lambda_t) + \alpha \langle A^\top \nabla f(A\lambda_t), \mathbf{e}_{j_{t+1}} \rangle$$

$$+ \frac{\alpha^2}{2} \sup_{\tau \in L_{t+1}} \mathbf{e}_{j_{t+1}}^\top A^\top \nabla^2 f(A(\lambda_t + \tau \mathbf{e}_{j_{t+1}})) A \mathbf{e}_{j_{t+1}}.$$

Next, using η from the definition of $\mathbb{G}$ granting $g'' \leq \eta g$ within the initial level set, and recalling that entries of A are within $[-1, 1]$, it follows that

$$\sup_{\tau \in L_{t+1}} \mathbf{e}_{j_{t+1}}^\top A^\top \nabla^2 f(A(\lambda_t + \tau \mathbf{e}_{j_{t+1}})) A \mathbf{e}_{j_{t+1}} = \sup_{\tau \in L_{t+1}} \sum_{i=1}^m g''(\mathbf{e}_i^\top A(\lambda_t + \tau \mathbf{e}_{j_{t+1}}) A_{ij_{t+1}}^2$$

$$\leq \eta \sup_{\tau \in L_{t+1}} \sum_{i=1}^m g(\mathbf{e}_i^\top A(\lambda_t + \tau \mathbf{e}_{j_{t+1}})$$

$$= \eta f(A\lambda_t),$$

the last step following by the definition of L_{t+1} and f. Inserting this second-order bound into the above Taylor expansion, for any $\alpha \in L_{t+1}$,

$$(f \circ A)(\lambda_t + \alpha \mathbf{e}_{j_{t+1}}) \leq f(A\lambda_t) + \alpha \langle A^\top \nabla f(A\lambda_t), \mathbf{e}_{j_{t+1}} \rangle + \frac{\alpha^2 \eta f(A\lambda_t)}{2}.$$

Observe that the lemma statement's choice of α_{t+1} is the minimizer of the convex quadratic of the right hand expression; plugging in α_{t+1} and rearranging yields the desired result. $\square$

16

G Deferred Material from Section 6

Proof of theorem 6.1. $((6.2) \implies (6.3))$ Let $\bar{\lambda} \in \mathbb{R}^n$ be given with $A\bar{\lambda} \in \mathbb{R}^m_{++}$, and let any increasing sequence $\{c_i\}_1^\infty \uparrow \infty$ be given. Then, since $f > \mathbf{0}_m$ and $\lim_{x \to \infty} g(x) = 0$,

$$\inf_\lambda f(A\lambda) \le \lim_{i \to \infty} f(c_i A\bar{\lambda}) = 0 \le \inf_\lambda f(A\lambda).$$

$((6.3) \implies (6.4))$ The point $\mathbf{0}_m$ is always dual feasible, and

$$\inf_\lambda f(A\lambda) = 0 = -f^*(-\mathbf{0}_m).$$

Since the dual optimum is unique (theorem 3.1), $\psi_A = \mathbf{0}_m$.

$((6.4) \implies (6.5))$ Suppose there exists $\psi \in \Phi_A$ with $\psi \ne \mathbf{0}_m$. Since $-f^*$ is continuous and increasing along every positive direction at $\mathbf{0}_m = \psi_A$ (see lemma C.2 and lemma C.3), there must exist some tiny $\tau > 0$ such that $-f^*(-\tau\psi) > -f^*(-\psi_A)$, contradicting the selection of ψ_A as a unique optimum.

$((6.5) \implies (6.2))$ This case is directly handled by Gordan's theorem (cf. theorem B.1). $\square$

H Deferred Material from Section 7

Attainability will be discussed rigorously in terms of *0-coercivity*—a function h is 0-coercive when every level set is compact (please see appendix A for further comments). The relevance to the current context is the following fact.

Proposition H.1. *Suppose f is differentiable, strictly convex, and $\mathrm{dom}(f) = \mathbb{R}^m$. Then $\inf_x f(x)$ is attainable iff f is 0-coercive.*

Note that while $g \in \mathbb{G}$ provides a strictly convex f, the function $f \circ A$ itself is not strictly convex. Instead, the strictly convex function $f + \iota_{\mathrm{Im}(A)}$ will be used when making statements about attainability.

Proof. It holds in general that 0-coercivity grants attainable minima (cf. [18, Proposition B.3.2.4] and [19, Proposition 1.1.3]), thus conversely let $\bar{x}$ be given with $f(\bar{x}) = \inf_x f(x)$. Consider any sublevel set $S_r := \{x \in \mathbb{R}^m : f(x) \le r\}$ with $r \ge f(\bar{x})$, and also any direction $d \in \mathbb{R}^m, \|d\|_2 = 1$. By strict convexity, for $t \ge 1$,

$$f(\bar{x} + td) > f(\bar{x} + d) + (t - 1)\langle \nabla f(\bar{x} + d), d\rangle.$$

Note by strict monotonicity of gradients (e.g. [18, B.4.1.4]) and first-order necessary conditions $(\nabla f(\bar{x}) = \mathbf{0}_m)$ that

$$\langle \nabla f(\bar{x} + d), d\rangle = \langle \nabla f(\bar{x} + d) - \nabla f(\bar{x}), \bar{x} + d - \bar{x}\rangle > 0,$$

and so, in every direction, the function eventually (i.e. for sufficiently large t) exceeds any r. As such, since the set of directions is compact, the level sets are bounded. That they are also compact follows since convex functions are closed on the interior of their domains. $\square$

Proof of theorem 7.1. $((7.2) \implies (7.3))$ As stated above, it suffices to show 0-coercivity of $f + \iota_{\mathrm{Im}(A)}$. Let $d \in \mathbb{R}^m \backslash \{\mathbf{0}_m\}$ and $\lambda \in \mathbb{R}^n$ be arbitrary. To show 0-coercivity, it suffices [18, Proposition B.3.2.4.iii] to show

$$\lim_{t \to \infty} \frac{f(A\lambda + td) + \iota_{\mathrm{Im}(A)}(A\lambda + td) - f(A\lambda)}{t} > 0. \tag{H.2}$$

If $d \notin \mathrm{Im}(A)$, then $\iota_{\mathrm{Im}(A)}(A\lambda + td) = \infty$. Suppose $d \in \mathrm{Im}(A)$; by (7.2), since $d \ne \mathbf{0}_m$, then $d \notin \mathbb{R}^m_+$ meaning there is at least one negative coordinate j. But then, since $g > 0$ and g is convex,

$$(\text{H.2}) \ge \lim_{t \to \infty} \frac{g(\mathbf{e}_j^\top(A\lambda + td)) - f(A\lambda)}{t} \ge \lim_{t \to \infty} \frac{g(\mathbf{e}_j^\top A\lambda) + td_j g'(\mathbf{e}_j^\top A\lambda) - f(A\lambda)}{t},$$

which is positive since $g'(\mathbf{e}_j^\top A\lambda) < 0$, and the other terms in the numerator vanish.

((7.3) $\implies$ (7.4)) Since the infimum is attainable, and thus designate any $\bar{\lambda}$ satisfying $\inf_\lambda f(A\lambda) = f(A\bar{\lambda})$ (note, although f is strictly convex, $f \circ A$ need not be, thus uniqueness is not guaranteed!). But then the conditions of [19, Exercise 3.3.9.f] are satisfied, meaning $\psi_A = -\nabla f(A\bar{\lambda})$, which is interior to $\mathbb{R}_+^m$ since $\nabla f \in \mathbb{R}_{--}^m$ everywhere (cf. lemma C.3).

((7.4) $\implies$ (7.5)) This holds since $\Phi_A \supseteq \{\psi_A\}$ and $\psi_A \in \mathbb{R}_{++}^m$.

((7.5) $\implies$ (7.2)) This case is directly handled by Stiemke's Theorem (cf. theorem B.4). $\qquad\square$

This section will now turn to the task of proving theorem 7.6. The lemmas will be stated with slightly more generality in order to allow their reuse in the proof of theorem 8.6.

First, the 0-coercivity property above immediately grants the existence of the tightest rectangle in the statement of theorem 7.6.

Lemma H.3. *Suppose $A \neq \mathbf{0}_{m\times n}$, $g \in \mathbb{G}$, and the infimum in (2.1) is attainable. Furthermore, let any $d \geq \inf_\lambda f(A\lambda)$ be given. Then there exists a (compact) tightest axis-aligned rectangle $\mathcal{C} \supseteq \{x \in \mathbb{R}^m : f(x) \leq d\}$. Furthermore, the dual image $-\nabla f(\mathcal{C}) \subset \mathbb{R}^m$ is also a compact axis-aligned rectangle, and moreover it is strictly contained within $-\mathrm{dom}(f^*) \subseteq \mathbb{R}_+^m$.*

Proof. Since $d \geq \inf_\lambda f(A\lambda)$, the level set $S_d := \{x \in \mathbb{R}^m : f(x) \leq d\}$ is nonempty. By proposition H.1, $f + \iota_{\mathrm{Im}(A)}$ is 0-coercive, meaning S_d is compact.

Now consider the rectangle $\mathcal{C}$ defined as a product of intervals $\mathcal{C} = \otimes_{i=1}^m [a_i, b_i]^m$, where

$$a_i := \inf\{x_i : x \in S_d\}, \qquad b_i := \sup\{x_i : x \in S_d\}.$$

By construction, $\mathcal{C} \supseteq S_d$, and furthermore any smaller axis-aligned rectangle must fail to include a piece of S_d. In particular, the tightest rectangle exists, and it is $\mathcal{C}$.

Finally, define $D := -\nabla f(\mathcal{C})$, the dual image of $\mathcal{C}$. Since ∇f is continuous, D is compact. Moreover, for any point $x \in \mathcal{C}$ and any open neighborhood $B(x, \delta)$ of x, the continuity of ∇f grants that $\nabla f(B(x, \delta)) \subseteq \mathrm{dom}(f^*)$, whereby it follows that every $\phi \in D$ is interior to $-\mathrm{dom}(f^*)$, and in particular D is contained within the interior of $-\mathrm{dom}(f^*)$.

Finally, note that $\nabla f(x) = (g'(x_1), g'(x_2), \ldots, g'(x_m))$, thus $D = -\otimes_{i=1}^m g'([a_i, b_i])$, an axis-aligned rectangle in the dual. $\qquad\square$

Next, dual distances are easily controlled along compact subsets of $\mathrm{Im}(A)$.

Lemma H.4. *Let $A \in \mathbb{R}^{m\times n}$, $g \in \mathbb{G}$, and any compact set S be given. Then f is strongly convex over S, and taking $c > 0$ to be the modulus of strong convexity, for any $x \in S \cap \mathrm{Im}(A)$,*

$$f(x) - \bar{f}_A \leq \frac{1}{2c} \inf_{\psi \in -\nabla f(S) \cap \mathrm{Ker}(A^\top)} \|\psi + \nabla f(x)\|_1^2.$$

Proof. Consider the optimization problem

$$\inf_{x\in S} \inf_{\substack{\phi\in\mathbb{R}^m \\ \|\phi\|_2=1}} \langle \nabla^2 f(x)\phi, \phi \rangle = \inf_{x\in S} \inf_{\substack{\phi\in\mathbb{R}^m \\ \|\phi\|_2=1}} \sum_{i=1}^m g''(x_i)\phi_i^2;$$

since S is compact and g'' is continuous, the infimum is attainable. But $g'' > 0$ and $\phi \neq \mathbf{0}_m$, meaning the infimum c is nonzero, and moreover it is the modulus of strong convexity of f over S [18, Theorem B.4.3.1.iii].

Now let any $x \in S \cap \mathrm{Im}(A)$ be given, and define $D = -\nabla f(S) \subset \mathbb{R}_+^m$. Consider the dual element $\mathsf{P}_{D\cap\mathrm{Ker}(A^\top)}^2(-\nabla f(x))$; due to the projection, it is dual feasible, and thus it must follow from theorem 3.1 that

$$\bar{f}_A = \sup\{-f^*(-\psi) : \psi \in \Phi_A\} \geq -f^*(-\mathsf{P}_{D\cap\mathrm{Ker}(A^\top)}^2(-\nabla f(x))).$$

Furthermore, since $x \in \mathrm{Im}(A)$,

$$\left\langle x, \mathsf{P}_{D\cap\mathrm{Ker}(A^\top)}^2(-\nabla f(x)) \right\rangle = 0.$$

18

Combining these with the Fenchel-Young inequality (cf. proposition B.10),

$$f(x) - \bar{f}_A \leq f(x) + f^*(-\mathsf{P}^2_{D \cap \mathrm{Ker}(A^\top)}(-\nabla f(x)))$$

$$= f^*(-\mathsf{P}^2_{D \cap \mathrm{Ker}(A^\top)}(-\nabla f(x))) + \langle \nabla f(x), x \rangle - f^*(\nabla f(x))$$

$$= f^*(-\mathsf{P}^2_{D \cap \mathrm{Ker}(A^\top)}(-\nabla f(x))) - f^*(\nabla f(x))$$

$$\quad - \left\langle \nabla f(x), -\mathsf{P}^2_{D \cap \mathrm{Ker}(A^\top)}(-\nabla f(x)) - \nabla f(x) \right\rangle$$

$$\leq \frac{1}{2c} \|\nabla f(x) + \mathsf{P}^2_{D \cap \mathrm{Ker}(A^\top)}(-\nabla f(x))\|_2^2,$$

where the last step follows by an application of lemma B.12, noting that both $\nabla f(x)$ and $-\mathsf{P}^2_{D \cap \mathrm{Ker}(A^\top)}(-\nabla f(x))$ are in $\nabla f(S) = -D$, and f is strongly convex with modulus c over S. To finish, rewrite P as an infimum and use $\|\cdot\|_2 \leq \|\cdot\|_1$. $\qquad\square$

Proof of theorem 7.6. Invoking lemma H.3 with $d = f(A\lambda_0)$ immediately provides a compact tightest axis-aligned rectangle $\mathcal{C}$ containing the initial level set $S := \{x \in \mathbb{R}^m : f(x) \leq f(A\lambda_0)\}$. Crucially, since the objective values never increase, S and $\mathcal{C}$ contain every iterate $A\lambda_t$.

Applying lemma H.4 to the set $\mathcal{C}$, then for any t,

$$f(A\lambda_t) - \bar{f}_A \leq \frac{1}{2c} \|\nabla f(A\lambda_t) + \mathsf{P}^2_{-\nabla f(\mathcal{C}) \cap \mathrm{Ker}(A^\top)}(-\nabla f(A\lambda_t))\|_1^2,$$

where $c > 0$ is the modulus of strong convexity of f over $\mathcal{C}$.

Finally, notice that $\mathcal{C}$ contains a minimizer, and thus $-\nabla f(\mathcal{C})$ contains the dual optimum ψ_A, which is dual feasible. Therefore $-\nabla f(\mathcal{C}) \cap \mathrm{Ker}(A^\top) \neq \emptyset$, and since lemma H.3 granted polyhedrality of $-\nabla f(\mathcal{C})$, proposition 4.3 provides $\gamma(A, -\nabla f(\mathcal{C})) > 0$. Plugging this into proposition 5.2 gives

$$f(A\lambda_{t+1}) - \bar{f}_A \leq f(A\lambda_t) - \bar{f}_A - \frac{\gamma(A, -\nabla f(\mathcal{C}))^2 \mathsf{D}^1_{-\nabla f(\mathcal{C}) \cap \mathrm{Ker}(A^\top)}(-\nabla f(A\lambda_t))^2}{2\eta f(A\lambda_t)}$$

$$\leq (f(A\lambda_t) - \bar{f}_A) \left(1 - \frac{c\gamma(A, -\nabla f(\mathcal{C}))^2}{\eta f(A\lambda_0)}\right),$$

and the result again follows by recursively applying this inequality. $\qquad\square$

I   Deferred Material from Section 8

Proof of theorem 8.1. ((8.2) $\implies$ (8.3)) Let $\bar{\lambda}$ be given with $A_0\bar{\lambda} \in \mathbb{R}^z_{++}$ and $A_+\bar{\lambda} = \mathbf{0}_p$, and let $\mathbb{R}_{++} \supset \{c_i\}_1^\infty \uparrow \infty$ be an arbitrary sequence increasing without bound. Lastly, let $\{\lambda_i\}_1^\infty$ be a minimizing sequence for $\inf_\lambda f(A_+\lambda)$. Then

$$\inf_\lambda f(A_+\lambda) = \lim_{i \to \infty} \left(f(A_+\lambda_i) + f(c_i A_0 \bar{\lambda})\right) \geq \inf_\lambda f(A\lambda) = \inf_\lambda (f(A_+\lambda) + f(A_0\lambda)) \geq \inf_\lambda f(A_+\lambda),$$

which used the fact that $f(A_0\lambda) \geq 0$ since $f \geq 0$. And since the chain of inequalities starts and ends the same, it must be a chain of equalities, which means $\inf_\lambda f(A_0\lambda) = 0$. To show attainability of $\inf_\lambda f(A_+\lambda)$, note the second part of (8.2) is one of the conditions of theorem 7.1.

((8.3) $\implies$ (8.4)) First, by theorem 6.1, $\inf_\lambda f(A_0\lambda) = 0$ means $\psi_{A_0} = \mathbf{0}_z$ and $\Phi_{A_0} = \{\mathbf{0}_z\}$. Thus

$$-f^*(-\psi_A) = \sup_{\psi \in \Phi_A} -f^*(-\psi)$$

$$= \sup_{\substack{\psi_z \in \mathbb{R}^z_+ \\ \psi_p \in \mathbb{R}^p_+ \\ A_0^\top \psi_z + A_+^\top \psi_p = \mathbf{0}_n}} -f^*(-\psi_z) - f^*(-\psi_p)$$

$$\geq \sup_{\psi_z \in \Phi_{A_0}} -f^*(-\psi_z) + \sup_{\psi_p \in \Phi_{A_+}} -f^*(-\psi_p)$$

$$= 0 - f^*(-\psi_{A_+}) = \inf_{\lambda \in \mathbb{R}^n} f(A_+\lambda) = \inf_{\lambda \in \mathbb{R}^n} f(A\lambda) = -f^*(-\psi_A).$$

Combining this with $g^*(0) = 0$ (cf. lemma C.2), $f^*(-\psi_A) = f^*(-\psi_{A_+}) = f^*(-\begin{bmatrix} \psi_{A_0} \\ \psi_{A_+} \end{bmatrix})$. But theorem 3.1 shows ψ_A was unique, which gives the result. And to obtain $\psi_{A_+} \in \mathbb{R}^p_{++}$, use theorem 7.1 with the 0-coercivity of $f + \iota_{\mathrm{Im}(A_+)}$.

((8.4) $\implies$ (8.5)) Since $\psi_{A_0} = \mathbf{0}_z$, it follows by theorem 6.1 that $\Phi_{A_0} = \{\mathbf{0}_z\}$. Furthermore, since $\psi_{A_+} \in \mathbb{R}^p_{++}$, it follows that $\Phi_{A_+} \cap \mathbb{R}^p_{++} \neq \emptyset$. Now suppose contradictorily that $\Phi_A \neq \Phi_{A_0} \times \Phi_{A_+}$; since it always holds that $\Phi_A \supseteq \Phi_{A_0} \times \Phi_{A_+}$, this supposition grants the existence of $\psi = \begin{bmatrix} \psi_z \\ \psi_p \end{bmatrix} \in \Phi_A$ where $\psi_z \in \mathbb{R}^z_+ \setminus \{\mathbf{0}_z\}$ and $\psi_p \in \mathbb{R}^p_+ \setminus \{\mathbf{0}_p\}$, meaning each have at least one positive coordinate. Now consider the element $q := \psi + \psi_A$ which has more nonzero elements than ψ_A, and is dual feasible. Setting A_q to just the rows of A corresponding to the nonzero entries of q, by theorem 7.1, the dual optimum ψ_{A_q} will have only nonzero entries. But this solution can be extended (by adding zeros) to an element $\psi'_{A_q} \in \Phi_A$, and moreover, since the selection of ψ_{A_q} did not choose ψ_A restricted to A_q (which would only remove zeros and thus maintain feasibility and dual objective value), it follows that $-f^*(-\psi'_{A_q}) > -f^*(-\psi_A)$, contradicting the definition of ψ_A as the dual optimum.

((8.5) $\implies$ (8.2)) Unwrapping the definition of Φ_A, the assumed statements imply

$$(\forall \phi_0 \in \mathbb{R}^z_+ \setminus \{\mathbf{0}_z\}, \phi_+ \in \mathbb{R}^p_+ \,\textbf{.}\, A_0^\top \phi_0 + A_+^\top \phi_+ \neq \mathbf{0}_n) \wedge (\exists \phi_+ \in \mathbb{R}^p_{++} \,\textbf{.}\, A_+^\top \phi_+ = \mathbf{0}_n).$$

Applying Motzkin's transposition theorem (cf. theorem B.7) to the left statement and Stiemke's theorem (cf. theorem B.4, which is implied by Motzkin's theorem) to the right yields

$$(\exists \lambda \in \mathbb{R}^n \,\textbf{.}\, A_0 \lambda \in \mathbb{R}^z_{++} \wedge A_+ \lambda \in \mathbb{R}^p_+) \wedge (\forall \lambda \in \mathbb{R}^n \,\textbf{.}\, A_+ \lambda \notin \mathbb{R}^p_+ \setminus \{\mathbf{0}_p\}),$$

which implies the desired statement. $\qquad\square$

Remark I.1. Consider the following iterative process. Start with all rows of A in A_+, and no rows in A_0. At every stage of this process, it will be maintained that A_0 satisfies the conditions of eq. (8.2), and when the process terminates, A_+ will satisfy its part of the statement. In the base case, the vector λ satisfying the guarantee for A_0 is the zero vector.

Iterate in the following manner. By Stiemke's theorem, either $A_+ \lambda \notin \mathbb{R}^m_+ \setminus \{\mathbf{0}_m\}$ for every λ, meaning we are done, or such a λ exists. If it exists, take the rows of $A_+ \lambda$ which were positive, and add move them to A_0; it must be shown that A_0 still satisfies the desired guarantee. Let λ' be the old choice; it will be shown that $\lambda + c\lambda'$ (for some $c > 0$) is a vector satisfying the desired guarantee on A_0. Consider any new row a_i moved to A_0, the old guarantee provides $\langle a_i, \lambda' \rangle = 0$, and thus $\langle a_i, \lambda' + \lambda \rangle > 0$ as needed. For any b_i that stayed in A_+, since this row was not moved, both $\langle b_i, \lambda \rangle = 0$ and the old guarantee gives $\langle b_i, \lambda' \rangle = 0$. What remains to be shown is that for any old row d_i of A_0, $\langle d_i, c\lambda' + \lambda \rangle > 0$. It may be the case that $\langle d_i, \lambda \rangle < 0$, but it is guaranteed that $\langle d_i, \lambda' \rangle > 0$; thus it suffices to choose a large $c > 0$ to dominate the influx of negative values.

This construction guarantees the existence of a satisfying decomposition, but note that it also guarantees uniqueness. If any other pair B_0, B_+ of matrices claimed to also satisfy the desired properties, then the vectors λ_A, λ_B satisfying the first conditioned could be added together, and the resulting vector λ' would violate the constraint on either A_+ or B_+. $\qquad\diamond$

Proof of theorem 8.6. Without loss of generality, suppose the entries of ψ_A are non-decreasing (and correspondingly permute the columns of A so that the A_0, A_+ stack together to form A. Let $A_0 \in \mathbb{R}^{z \times n}$ and $A_+ \in \mathbb{R}^{p \times n}$ be the matrices corresponding to the zero and positive entries of ψ_A (thus A is A_0 on top of A_+, with $m = z + p$). By theorem 8.1, $\bar{f}_{A_+} = \bar{f}_A$, and the form of f gives $f(A\lambda_t) = f(A_0 \lambda_t) + f(A_+ \lambda_t)$, thus

$$f(A\lambda_t) - \bar{f}_A = f(A_0\lambda_t) + f(A_+\lambda_t) - \bar{f}_{A_+}. \tag{I.2}$$

For the left term, since $g(x) \leq \beta |g'(x)|$,

$$f(A_0\lambda_t) \leq \beta \|\nabla f(A_0\lambda_t)\|_1 = \beta \|\nabla f(A_0\lambda_t) + \mathsf{P}^1_{\Phi_{A_0}}(\nabla f(A_0\lambda_t))\|_1,$$

which used the fact (from theorem 8.1) that $\Phi_{A_0} = \{\mathbf{0}_z\}$.

For the right term of (I.2), recall from theorem 8.1 that $f + \iota_{\mathrm{Im}(A_+)}$ has attainable minima and is thus 0-coercive by proposition H.1; as such, the level set $S_+ := \{x \in \mathbb{R}^p : f(x) \le f(A\lambda_0)\}$ is compact. Note that, for all t, $A_+\lambda_t \in S_+$. This follows from

$$f(A\lambda_0) \ge f(A\lambda_t) = f(A_0\lambda_t) + f(A_+\lambda_t) \ge f(A_+\lambda_t),$$

which used $f \ge 0$. It is crucial that the level set compares against $f(A\lambda_0)$ and not $f(A_+\lambda_0)$.

Continuing, lemma H.3 may be applied to A_+ with value $d = f(A\lambda_0)$, granting a tightest axis-aligned rectangle $\mathcal{C}_+ \subseteq \mathbb{R}_+^p$ containing S_+. Applying lemma H.4 to A_+ and $\mathcal{C}_+$, f is strongly convex with modulus $c > 0$ over $\mathcal{C}_+$, and for any t,

$$f(A_+\lambda_t) - \bar{f}_{A_+} \le \frac{1}{2c}\|\nabla f(A_+\lambda_t) + \mathsf{P}^1_{-\nabla f(\mathcal{C}_+) \cap \mathrm{Ker}(A^\top)}(-\nabla f(A_+\lambda_t))\|_1^2.$$

Next, set $w := \sup_t \|\nabla f(A_+\lambda_t) + \mathsf{P}^2_{-\nabla f(\mathcal{C}_+) \cap \mathrm{Ker}(A^\top)}(-\nabla f(A_+\lambda_t))\|_1$; $w < \infty$ since S_+ is compact, and $-\nabla f(\mathcal{C}_+) \cap \mathrm{Ker}(A^\top)$ is nonempty. By definition of w,

$$\|\nabla f(A_+\lambda_t) + \mathsf{P}^1_{\Phi_{A_+}}(-\nabla f(A_+\lambda_t))\|_1^2 \le w\|\nabla f(A_+\lambda_t) + \mathsf{P}^1_{\Phi_{A_+}}(-\nabla f(A_+\lambda_t))\|_1.$$

Now to combine the upper bounds for the left and right terms of (I.2). The principal claim is that, for any $\phi = \phi_z \times \phi_p \in \mathbb{R}^m$,

$$\mathsf{P}^1_{(\mathbb{R}_+^z \times -\nabla f(\mathcal{C}_+)) \cap \mathrm{Ker}(A^\top)}\left(\begin{bmatrix}\phi_z \\ \phi_p\end{bmatrix}\right) = \begin{bmatrix}\mathbf{0}_z \\ \mathsf{P}^1_{-\nabla f(\mathcal{C}_+) \cap \mathrm{Ker}(A^\top)}(\phi_p)\end{bmatrix} = \begin{bmatrix}\mathsf{P}^1_{\Phi_{A_0}}(\phi_z) \\ \mathsf{P}^1_{-\nabla f(\mathcal{C}_+) \cap \mathrm{Ker}(A^\top)}(\phi_p)\end{bmatrix}.$$

First notice $\mathbb{R}_+^m \times -\nabla f(\mathcal{C}_+) \subseteq \mathbb{R}_+^m$, thus $(\mathbb{R}_+^m \times -\nabla f(\mathcal{C}_+)) \cap \mathrm{Ker}(A^\top) \subseteq \Phi_A$. Recall from theorem 8.1 that $\Phi_A = \Phi_{A_0} \times \Phi_{A_+} = \{\mathbf{0}_z\} \times \Phi_{A_+}$, and $\Phi_{A_0} = \{\mathbf{0}_z\}$, from which it follows that the final latter two quantities above are valid projections. To finish, notice that if there were a closer projection, it would grant a closer projection to $-\nabla f(\mathcal{C}_+) \cap \mathrm{Ker}(A_+^\top)$, a contradiction, and the result follows. (When the projections are not unique, the discussed choice is still a valid choice; moreover, it matters not in the following distance computations.)

Thus

$$\begin{aligned}(\mathrm{I.2}) &\le \beta\|\nabla f(A_0\lambda_t) + \mathsf{P}^1_{\Phi_{A_0}}(-\nabla f(A_0\lambda_t))\|_1 \\ &\quad + w(2c)^{-1}\|\nabla f(A_+\lambda_t) + \mathsf{P}^1_{-\nabla f(\mathcal{C}_+) \cap \mathrm{Ker}(A^\top)}(-\nabla f(A_+\lambda_t))\|_1 \\ &\le (\beta + w(2c)^{-1})\|\nabla f(A\lambda_t) + \mathsf{P}^1_{(\mathbb{R}_+^z \times -\nabla f(\mathcal{C}_+)) \cap \mathrm{Ker}(A^\top)}(\nabla f(A\lambda_t))\|_1.\end{aligned}$$

Note that $\mathbb{R}_+^z \times -\nabla f(\mathcal{C}_+)$ is polyhedral, and $\mathcal{C}_+$ contains a minimizer for $f \circ A_+$, meaning $\mathbb{R}_+^z \times -\nabla f(\mathcal{C}_+)$ contains the dual feasible point $\mathbf{0}_z \times \psi_{A_+}$, thus proposition 4.3 gives $\gamma(A, \mathbb{R}_+^z \times -\nabla f(\mathcal{C}_+)) > 0$. By an application of proposition 5.2 and making use of $f(A\lambda_t) \le f(A\lambda_0)$,

$$\begin{aligned}f(A\lambda_{t+1}) - \bar{f}_A &\le f(A\lambda_t) - \bar{f}_A - \frac{\gamma(A, \mathbb{R}_+^z \times -\nabla f(\mathcal{C}_+))^2 \mathsf{D}^1_{(\mathbb{R}_+^z \times -\nabla f(\mathcal{C}_+)) \cap \mathrm{Ker}(A^\top)}(-\nabla f(A\lambda_t))^2}{2\eta f(A\lambda_t)} \\ &\le f(A\lambda_t) - \bar{f}_A - \frac{\gamma(A, \mathbb{R}_+^z \times -\nabla f(\mathcal{C}_+))^2 (f(A\lambda_t) - \bar{f}_A)^2}{2(\beta + w/(2c))^2 \eta f(A\lambda_0)}.\end{aligned}$$

Now apply lemma B.11 with $r := \min\left\{1, \frac{\gamma(A, \mathbb{R}_+^z \times -\nabla f(\mathcal{C}_+))^2}{(\beta + w/(2c))^2 \eta}\right\}/(2f(A\lambda_0))$. $\qquad\square$

Proof of theorem 8.7. This proof proceeds in two stages: first the gap between any solution with l^1 norm B is shown to be large, and then it is shown that the l^1 norm of the BOOST solution (under logistic loss) grows slowly.

To start, $\mathrm{Ker}(S^\top) = \{z(1, 1, 0) : z \in \mathbb{R}\}$, and $-g^*$ is maximized at $g'(0)$ with value $-g(0)$ (cf. lemma C.2). Thus $\psi_S = (-g'(0), -g'(0), 0)$, and $\bar{f}_S = -f^*(-\psi_S) = 2g(0) = 2\ln(2)$.

Next, by calculus, given any B,

$$\begin{aligned}\inf_{\|\lambda\|_1 \le B} f(S\lambda) - \bar{f}_S &= f\left(S\begin{bmatrix}B/2 \\ B/2\end{bmatrix}\right) - 2\ln(2) \\ &= (2\ln(2) + \ln(1 + \exp(-B))) - 2\ln(2) \\ &= \ln(1 + \exp(-B)).\end{aligned}$$

Now to bound the l^1 norm of the iterates. By the nature of exact line search, the coordinates of λ are updated in alternation (with arbitrary initial choice); thus let u_t denote the value of the coordinate updated in iteration t, and v_t be the one which is held fixed. (In particular, $v_t = u_{t-1}$.)

The objective function, written in terms of (u_t, v_t), is

$$\ln(1 + \exp(v_t - u_t)) + \ln(1 + \exp(u_t - v_t)) + \ln(1 + \exp(-u_t - v_t)).$$

Differentiating and collecting terms, u_t and v_t must satisfy, after the line search

$$\exp(2u_t) = \exp(2v_t) + 2\exp(v_t - u_t) + 2. \tag{I.3}$$

First it will be shown for $t > 1$, by induction, that $u_t \geq v_t$. The base case follows by inspection (since $u_0 = v_0 = 0$ and so $u_1 = \ln(2)$). Now the inductive hypothesis grants $u_t \geq v_t$; the case $u_t = v_t$ can be directly handled by eq. (I.3), thus suppose $u_t > v_t$. But previously, it was shown that the optimal l^1 bounded choice has both coordinates equal; as such, the current iterate, with coordinates (u_t, v_t), is worse than the iterate (u_t, u_t), and thus the line search will move in a positive direction, giving $u_{t+1} \geq v_{t+1}$.

It will now be shown by induction that, for $t \geq 1$, $u_t \leq \frac{1}{2}\ln(4t)$. The base case follows by the direct inspection above. Applying the inductive hypothesis to the update rule above, and recalling $v_{t+1} = u_t$ and that the weights increase (i.e., $u_{t+1} \geq v_{t+1} = u_t$),

$$\exp(2u_{t+1}) = \exp(2u_t) + 2\exp(u_t - u_{t+1}) + 2 \leq \exp(2u_t) + 2\exp(u_t - u_t) + 2 \leq 4t + 4 \leq 4(t+1).$$

To finish, recall by Taylor expansion that $\ln(1 + q) \geq q - \frac{q^2}{2}$; consequently for $t \geq 1$

$$f(S\lambda_t) - \bar{f}_S \geq \inf_{\|\lambda\|_1 \leq \ln(4t)} f(S\lambda) - \bar{f}_S \geq \ln\left(1 + \frac{1}{4t}\right) \geq \frac{1}{4t} - \frac{1}{2}\left(\frac{1}{4t}\right)^2 \geq \frac{1}{8t}. \qquad \square$$