Steepest Descent Analysis for Unregularized Linear Prediction with Strictly Convex Penalties

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 presents a convergence analysis, generalized from a study of boosting [1], of unregularized linear prediction. Here the empirical risk — incorporating strictly convex penalties composed with a linear term — may fail to be strongly convex, or even attain a minimizer. This analysis is demonstrated on linear regression, decomposable objectives, and boosting.

1 Introduction

Consider any linear prediction problem, where the optimization variable $\lambda \in \mathbb{R}^n$ interacts with training data accumulated row-wise into a matrix $A \in \mathbb{R}^{m\times n}$, and a good fit is achieved by approximately minimizing the objective

$$\inf\left\{f(A\lambda) : \lambda \in \mathbb{R}^n\right\}, \tag{1.1}$$

where $f : \mathbb{R}^m \to \mathbb{R}$ is strictly convex, continuously differentiable, bounded below, and finite everywhere. This formulation has attracted interest both within the regression community, where f often penalizes the squared l^2 distance to some target observations, and within the classification community, the flagship application being boosting, where f is a convex surrogate to the 0/1 loss.

The goal of this manuscript is to provide a convergence analysis of (1.1) as minimized by steepest descent with line search, with sensitivity to two issues which complicated earlier analyses: the situation that A has deficient column rank, and the case that $f \circ A$ lacks minimizers.

To review the primary known results, suppose A has full column rank and f is 0-coercive (level sets are compact); then over the initial level set, the (assumed extant) Hessian $A^\top \nabla^2 f(\cdot) A$ is positive definite, and the convergence rate is $\mathcal{O}(\ln(1/\epsilon))$, meaning this number of iterations suffices to attain accuracy $\epsilon > 0$ [2, (9.18)]. If A is an arbitrary matrix, but supposing $f \circ A$ has minimizers, then steepest descent of (1.1) is known to exhibit *Q-linear convergence*, meaning a rate $\mathcal{O}(\ln(1/\epsilon))$ if one ignores some finite prefix of the iterate sequence [3].

The present manuscript will establish an immediate rate of $\mathcal{O}(\ln(1/\epsilon))$ when $f \circ A$ has minimizers (cf. Corollary 2.5); more importantly, it will also outline conditions granting $\mathcal{O}(\ln(1/\epsilon))$ or $\mathcal{O}(1/\epsilon)$, but without relying on the presence of minimizers (cf. Theorem 2.4). The analysis is generalized from a study of boosting, where it was used to drop the convergence rate under a family of losses (including the exponential and logistic losses) from $\mathcal{O}(\exp(1/\epsilon^2))$ to $\mathcal{O}(1/\epsilon)$, and in some cases $\mathcal{O}(\ln(1/\epsilon))$ [1].

Remark 1.2. Although the focus here is to analyze a specific algorithm, note briefly how certain more contemporary methods fare on (1.1). Under basic structural assumptions (e.g., Lipschitz gradients), if $f \circ A$ has minimizers but A is an arbitrary matrix, the methods of mirror descent, as well as various versions of Nesterov's accelerated gradient methods, will respectively attain the rates

$\mathcal{O}(1/\epsilon)$ and $\mathcal{O}(1/\sqrt{\epsilon})$ [4, 5]. Due to numerical and statistical considerations, one would typically regularize (1.1) in some way, and these rates would form the general case.

Hidden within the $\mathcal{O}(\cdot)$ of the rates for these methods is a Bregman divergence $D(\hat{\lambda}, \lambda_0)$ between some user-selected reference point $\hat{\lambda}$, and the initial iterate λ_0. When $f \circ A$ has a minimizer, it may be used as the reference point $\hat{\lambda}$, and $D(\hat{\lambda}, \lambda_0)$ can be safely treated as a constant. But when $f \circ A$ lacks minimizers (i.e., no regularization was added, and one is again considering (1.1) verbatim), these suboptimality bounds may be instantiated with elements of a family of increasingly optimal reference points $\{\hat{\lambda}_\epsilon\}_{\epsilon \downarrow 0}$. Since $D(\cdot, \cdot)$ is generally chosen to upper bound some squared norm (e.g., see Tseng [5, eq. (7)]), unattainability of the infimum implies $D(\hat{\lambda}_\epsilon, \lambda_0) \uparrow \infty$ as $\epsilon \downarrow 0$. Said another way, this Bregman term is no longer a constant, but instead encodes a dependence on a chosen suboptimality ϵ. As a result, the asymptotics of these rates become unclear (what if $D(\hat{\lambda}_\epsilon, \lambda_0) \geq 1/\epsilon$ for every ϵ?). On the other hand, the analysis here shows that this tricky dependence need not be present at least in the case of steepest descent with line search. $\lozenge$

2 Convergence Analysis

2.1 Background

For concreteness, the following definition of steepest descent is adopted. Starting with $t = 1$ and some provided iterate $\lambda_0 \in \mathbb{R}^n$, the following steps are repeated ceaselessly.

1. Choose steepest descent direction v_t with respect to $\|\cdot\|$:
$$v_t \in \operatorname{Arg\,min} \left\{ \langle v, \nabla(f \circ A)(\lambda_{t-1}) \rangle : \|v\| \leq 1 \right\}.$$
Note $\langle v_t, \nabla(f \circ A)(\lambda_{t-1}) \rangle = -\|\nabla(f \circ A)(\lambda_{t-1})\|_*$.
2. Choose step size α_t via line search (i.e., approximately minimize $\alpha \mapsto (f \circ A)(\lambda_{t-1} + \alpha v_t)$).
3. Update $\lambda_t := \lambda_{t-1} + \alpha_t v_t$, then $t := t + 1$.

Popular choices for $\|\cdot\|$ are gradient descent ($\|\cdot\|_2$) and coordinate descent ($\|\cdot\|_1$).

The analysis here is based upon two very simple observations regarding the convex dual

$$(1.1) = \sup \left\{ -f^*(\phi) : \phi \in \ker(A^\top) \right\}; \tag{2.1}$$

specifically, the influence of the problematic matrix A is moved out of the objective, and the dual optimum is attainable and unique (these facts follow from Fenchel duality and properties of f [6, 1, Theorem 3.3.5, Theorem 4]). Recalling that convergence analyses typically proceed by controlling the distance to an optimum, the approach here is to simply map the problem into the dual, and to then proceed as usual.

Executing this map takes two steps. First, standard line search guarantees relate the single-step suboptimality to a gradient term $\|\nabla(f \circ A)(\lambda_t)\|_* = \|A^\top \nabla f(A\lambda_t)\|_*$. The second step is to replace this with a suboptimality notion in the dual, which will use the following quantity.

Definition 2.2. For any closed convex nonempty set C, define $\mathsf{P}_C^{\|\cdot\|}(x) \in \operatorname{Arg\,min}_{y \in C} \|y - x\|$, with an arbitrary choice made in the presence of nonuniqueness. Let any polyhedron $S \subseteq \mathbb{R}^m$ with $S \setminus \ker(A^\top) \neq \emptyset$ and $S \cap \ker(A^\top) \neq \emptyset$ be given. Fixing two norms $\|\cdot\|$ and $|||\cdot|||$, define

$$\gamma(A, S) := \inf \left\{ \frac{\|A^\top \phi\|_*}{|||\phi - \mathsf{P}_{S \cap \ker(A^\top)}^{|||\cdot|||_*}(\phi)|||_*} : \phi \in S \setminus \ker(A^\top) \right\}.$$

Crucially, $\gamma(A, S) > 0$. (It suffices to apply equivalence of norms on (finite dimensional!) Euclidean space to the l^1/l^∞ version of $\gamma(A, S)$ [1, Theorem 9].) $\lozenge$

This quantity is simply a lower bound on the ratio between the normed gradient term $\|A^\top \nabla f(A\lambda_t)\|_*$, and the distance from the dual iterate $\nabla f(A\lambda_t)$ to a restriction $S \cap \ker(A^\top)$ of the dual feasible set. Although initially exotic, $\gamma(A, S)$ can be derived from the weak learning rate in boosting [1, Appendix F]. In the original presentation, a *weak learning assumption*, a structural property on A, was needed to grant positivity [7]; the definition here foregoes that need.

2.2 Main Result

The convergence results will depend on three structural properties, labeled (A)-(C).

(A) f is strictly convex, continuously differentiable, bounded below, and everywhere finite.

A lower bound grants that the infimum is finite, and finiteness means the only constraints involved are the implicit affine constraints imposed by A. the other two parts will be discussed with (C).

(B) Gradients $\nabla(f \circ A)$ are Lipschitz continuous with constant L_t with respect to norm $\|\cdot\|$ for any λ, λ' with $\max\{f(A\lambda), f(A\lambda')\} \leq f(A\lambda_t)$:

$$\|\nabla(f \circ A)(\lambda) - \nabla(f \circ A)(\lambda')\|_* \leq L_t \|\lambda - \lambda'\|.$$

Lipschitz gradients are an easy way to provide a nice line search guarantee (cf. the proof sketch of Theorem 2.4). While it may seem unusual to specialize this bound for every level set, this refinement is key when proving $\mathcal{O}(\ln(1/\epsilon))$ rates for boosting under the weak learning assumption (cf. Example 3.5).

(C) There is a polyhedron S containing the dual optimum, and every dual iterate ($S \supseteq \{\nabla f(A\lambda_t)\}_0^\infty$). Furthermore, for some norm $|||\cdot|||$, a scalar $C_d > 0$, and for all t, taking $\phi_t := \nabla f(A\lambda_t)$ for convenience,

$$C_d \left(\inf_{\phi \in S \cap \ker(A^\top)} f^*(\phi) - f^*(\phi_t) - \langle \nabla f^*(\phi_t), \phi - \phi_t \rangle \right)^k \leq \frac{|||\phi_t - \mathsf{P}^{|||\cdot|||_*}_{S \cap \ker(A^\top)}(\phi_t)|||_*^2}{L_t}$$

for some $k \in \{1, 2\}$.

To demystify this expression, first notice that the infimand is the f^*-Bregman divergence from ϕ_t to the closest ϕ within a restriction of the dual feasible set. The other two conditions of (A) now come into play: the strict convexity and differentiability of f respectively grant differentiability and strict convexity of f^* [8, Section E.4.1], which are sufficient for this expression to be well-defined and nonzero.

In the case of strong convexity, (C) may be interpreted more familiarly. Suppose S is compact and interior to $\mathrm{dom}(f^*)$, whereby the primal image $\nabla f^*(S)$ is compact (cf. (A) and Hiriart-Urruty and Lemaréchal [8, E.4.1.1]). Recall the following result [9, Lemma 18]: when f is strongly convex with modulus c over $\nabla f^*(S)$ with respect to norm $|||\cdot|||$, and any $\phi, \phi' \in S$ are given,

$$f^*(\phi) - f^*(\phi') - \langle \nabla f^*(\phi'), \phi - \phi' \rangle \leq \frac{1}{2c} |||\, \phi - \phi' \,|||_*^2 . \tag{2.3}$$

As such, making the simplifying choice $L_t := L_1$ for all t, then (C) is satisfied with $k = 1$ simply by setting $C_d = 2c/L_1$, and instantiating (2.3) with $\phi' = \phi_t$, $\phi = \mathsf{P}^{|||\cdot|||_*}_{S \cap \ker(A^\top)}(\phi_t)$, the latter value being considered in the infimum within (C). Although (C) may appear baroque in the presence of strong convexity, the extra parts are beneficial in its absence.

With the help of these properties, the convergence result may finally be stated.

Theorem 2.4. *Let f, A be given, and suppose (A) and (B) hold.*

- *If (C) is satisfied with $k = 2$, then the rate of convergence is $\mathcal{O}(1/\epsilon)$.*
- *If (C) is satisfied with $k = 1$, then the rate of convergence is $\mathcal{O}(\ln(1/\epsilon))$.*

As will be demonstrated in Example 3.1, the quantities hidden by the $\mathcal{O}(\cdot)$ can be recovered by inspecting the proof of Theorem 2.4 (specifically (4.4)).

Corollary 2.5. *Suppose f, A satisfy properties (A) and (B). If $f \circ A$ has minimizers, then (C) can be satisfied with $k = 1$, granting a rate $\mathcal{O}(\ln(1/\epsilon))$.*

3 Examples

Example 3.1 (Quadratics)**.** For the sake of illustration, consider a convex quadratic $\frac{1}{2}\lambda^\top Q\lambda$ as solved by gradient descent (i.e., choose $\|\cdot\|$ to be $\|\cdot\|_2$). When Q is symmetric positive definite,

$$\lambda_t^\top Q\lambda_t \leq (1-\sigma_{\text{rank}(Q)}(Q)/\sigma_{\max}(Q))^t(\lambda_0^\top Q\lambda_0), \tag{3.2}$$

meaning a rate $\mathcal{O}(\ln(1/\epsilon))$ (see for instance (9.18) in Boyd and Vandenberghe [2], since $\sigma_{\text{rank}(Q)}(Q)I \preceq Q \preceq \sigma_{\max}(Q)I$).

When Q is merely symmetric positive semi-definite, the above analysis fails, but this example within the present framework with $f = \frac{1}{2}\|\cdot\|_2^2$ and $A := \sqrt{Q}$ (as defined by the spectral decomposition of Q; note $\sigma_i(A) = \sqrt{\sigma_i(A^\top A)} = \sqrt{\sigma_i(Q)}$). As stated in the introduction, a result of Luo and Tseng [3] grants Q-linear convergence. Meanwhile, $\nabla(f\circ A)(\cdot) = A^\top A(\cdot)$ is Lipschitz with uniform constant $L_t = \sigma_{\max}(Q)$, thus (A) and (B) hold, and Corollary 2.5 may be applied, granting an immediate rate $\mathcal{O}(\ln(1/\epsilon))$.

Inspecting (4.4) in the proof of Theorem 2.4 leads to an analog of (3.2). Specifically, choosing polyhedron $S := \mathbb{R}^m$ (which certainly contains all dual iterates and the dual optimum), strong convexity of f with respect to the norm $|||\cdot||| = \|\cdot\|_2$ grants, via (2.3), that (C) holds with $C_1 = 2/L_t = 2/\sigma_{\max}(Q)$. Furthermore, $\gamma(A,\mathbb{R}^m)$ may be explicitly computed: it is simply $\sqrt{\sigma_{\text{rank}(Q)}(Q)}$. Instantiating (4.4) with these terms gives

$$\lambda_t^\top Q\lambda_t \leq (1-\sigma_{\text{rank}(Q)}(Q)/(3\sigma_{\max}(Q)))^t(\lambda_0^\top Q\lambda_0),$$

where the 3 is due to the choice of approximate line search providing (4.4). ♢

Example 3.3 (Linear regression)**.** Consider first the problem of least squares linear regression, where the training examples $(x_i)_{i=1}^m$ are collected row-wise into the matrix A ($(y_i)_{i=1}^m$ are rolled into f_{ls}):

$$\inf_\lambda f_{\text{ls}}(A\lambda) = \inf_\lambda \frac{1}{2}\|Y-A\lambda\|_2^2 = \inf_\lambda \frac{1}{2}\sum_{i=1}^m (y_i - \langle x_i,\lambda\rangle)^2.$$

The optimization behavior is close to Example 3.1's convex quadratic: a result of [3] provides Q-linear convergence, and Corollary 2.5 here provides a rate $\mathcal{O}(\ln(1/\epsilon))$. Of course, given the advanced state of linear least squares solvers, it may seem silly to apply a black box descent method to this problem. As such, consider a different objective function, for instance one aiming for robustness (not penalizing large errors too much). One choice is the Huber loss, which although not strictly convex, has a nice strictly convex approximant:

$$f_{\text{lc}}(A\lambda) = \sum_{i=1}^m \ln\left(\cosh(y_i - \langle x_i,\lambda\rangle)\right).$$

This is strongly convex within the initial level set, and Corollary 2.5 again grants $\mathcal{O}(\ln(1/\epsilon))$. Note however that the modulus of strong convexity (which controls C_d here) will be minuscule, and inspecting (4.4), the hidden terms in the rate will be correspondingly massive.

The above discussion left open the question of descent method. Gradient descent (i.e., $\|\cdot\| = \|\cdot\|_2$) will lead to dense solutions, so consider coordinate descent (i.e., $\|\cdot\| = \|\cdot\|_1$), which heuristically provides sparsity. Although the convergence rate is still $\mathcal{O}(\ln(1/\epsilon))$, the hidden factors smash any hope of actual sparsity. A more refined analysis (with respect to sparsity) is to compare to a certain sparse predictor, as in the results of Shalev-Shwartz, Srebro, and Zhang [10]; perhaps there is some way to combine those sparsity-sensitive results — which do not separate the impact of A — with the fast convergence here. ♢

Example 3.4 (Decomposable objectives)**.** Suppose establishing the second order bound (C) seems insurmountable for some matrix A, but A can be broken into pieces A_1, A_2 so that:

- There is a primal decomposition: for any λ_t,

$$f(A\lambda_t) - \inf_\lambda f(A\lambda) \leq \sum_{j\in\{1,2\}} f(A_j\lambda_t) - \inf_\lambda f(A_j\lambda).$$

- There are independent bounds in the dual: (C) holds for each A_j with $k = 2$ and some S_j.
- There is a dual decomposition: for some S, $c > 0$, and any $\phi_t := \nabla f(A\lambda_t)$,

$$\sum_{j\in\{1,2\}} |||\phi_t - \mathsf{P}^{|||\cdot|||_*}_{S_j\cap\ker(A_j^\top)}(\phi_t)|||_* \leq c \,|||\, \phi_t - \mathsf{P}^{|||\cdot|||_*}_{S\cap\ker(A^\top)}(\phi_t)\,|||_* \,.$$

After some algebra, making use of (4.3) in the proof of Theorem 2.4, (C) holds for the full problem with $k = 2$, granting a rate $\mathcal{O}(1/\epsilon)$. As an example application, seeing how attainability simplifies the problem, consider splitting $\mathbb{R}^n$ into two orthogonal subspaces: the linear hull of all directions λ satisfying $(f \circ A)'_\infty(\lambda) := \lim_{t\to\infty}(f(tA\lambda) - f(\mathbf{0}_m))/t = 0$, and the orthogonal complement of this subspace. This function $(f\circ A)'_\infty$, the *asymptotic function of* $f\circ A$, is closed and convex, and the resulting subspace is relatively easy to characterize [8, Proposition B.3.2.4], and one can compose A with the projection onto each subspace to obtain the above $\{A_j\}_{1,2}$. The problem now reduces to controlling the infinite piece, feeling content with the finite piece (thanks to Corollary 2.5), and producing the above decomposition. Exactly this strategy was followed in order to analyze boosting [1], as will be sketched in Example 3.5.

The details of this decomposition and how the pieces combine may vary, but the general approach of splitting along an orthogonal subspace pair is widely applicable. Interestingly, such a decomposition was used by Agarwal, Negahban, and Wainwright [11] to circumvent a reliance upon strong convexity in high dimensional problems. ♢

Example 3.5 (Boosting [1])**.** Now consider the case of boosting the accuracy of a class of binary weak predictors. Since there are m examples, there are at most 2^m distinct weak predictors (i.e. they can be finitely indexed), and thus set $A_{ij} := -y_i h_j(x_i)$. Classical boosting [7] minimizes

$$f_{\mathrm{b}}(A\lambda) = \sum_{i=1}^m g_{\mathrm{b}}(\mathbf{e}_i^\top A\lambda),$$

via coordinate descent, where $g_{\mathrm{b}}(x) \geq \mathbb{1}[x \geq 0]$ (or some scaling thereof). Typically $\lim_{x\to-\infty} g_{\mathrm{b}}(x) = 0$, which combined with strict convexity implies minimizers may fail to exist, and standard descent analyses do not apply.

Under some regularity conditions on g_{b} (which hold for the exponential and logistic losses), a rate of $\mathcal{O}(1/\epsilon)$ can be shown via the decomposition strategy of Example 3.4 [1]. In the parlance of boosting, the matrix A is decomposed into those rows a_i where every minimizing sequence leads to infinite dot products $\langle a_i, \lambda\rangle$ (i.e., the *margins* grow unboundedly), and those where they do not. On those where they stay bounded, as discussed in Example 3.5, strong convexity is easily exhibited, granting (C) on that subproblem. For the other rows, the aforementioned regularity conditions on g_{b} encode a *flattening condition*: gradients cannot become too flat without objective values becoming tiny, and again (C) follows (cf. [1, discussion after proof of Theorem 27]).

Interestingly, if all rows fall into either case, then the stronger guarantee of Theorem 2.4 may be applied, granting rate $\mathcal{O}(\ln(1/\epsilon))$. In fact, in the purely unbounded margin case, the proof can be seen as an elaborate reworking of the original AdaBoost convergence rate under the weak learning assumption [7]. It is precisely in this case that exploiting the denominator L_t in the definition of (C) is necessary to exhibit the faster rate (i.e., to establish the bound with $k = 1$). ♢

4 Proof Sketches

Proof of Theorem 2.4. Pieces of this proof can be found throughout an analysis of boosting [1]. To start, the line search, combined with (B), provides a single-iteration improvement of

$$f(A(\lambda_t + \alpha_t v_t)) \leq f(A\lambda_t) - \frac{\|A^\top\nabla(A\lambda_t)\|_*^2}{6L_t}. \tag{4.1}$$

(The $1/3$ is an artifact of using a Wolfe search with $c_1 = 1/3$ and $c_2 = 1/2$ [12, 1, Lemma 3.1, Proposition 38]; other line searches provide similar guarantees.)

Next, if λ_t is optimal, there is nothing to do, so suppose it is suboptimal. Thus $\nabla f(A\lambda_t) \in S \setminus \ker(A^\top)$, so the infimum of $\gamma := \gamma(A, S)$ may be instantiated with $\nabla f(A\lambda_t)$, meaning

$$\|A^\top\nabla f(A\lambda_t)\|_* \geq \gamma \,|||\, \nabla f(A\lambda_t) - \mathsf{P}^{|||\cdot|||_*}_{S\cap\ker(A^\top)}(\nabla f(A\lambda_t))\,|||_* \,.$$

Plugging this into (4.1),

$$f(A(\lambda_t + \alpha_t v_t)) \leq f(A\lambda_t) - \frac{\gamma^2 \left\| \left\| \nabla f(A\lambda_t) - \mathsf{P}^{\|\|\cdot\|\|_*}_{S\cap\ker(A^\top)}(\nabla f(A\lambda_t)) \right\|\right\|_*^2}{6L_t}. \tag{4.2}$$

Next, by: the Fenchel-Young inequality, (2.1), $A^\top \mathsf{P}^{\|\|\cdot\|\|_*}_{S\cap\ker(A^\top)}(\nabla f(A\lambda_t)) = 0$, $\nabla f^*(\nabla f(A\lambda)) = A\lambda$, S containing the dual optimum, and setting $\phi_t := \nabla f(A\lambda_t)$ for convenience,

$$\begin{aligned} f(A\lambda_t) - \inf_\lambda f(A\lambda) &= \inf\left\{ f^*(\phi_t) - f^*(\phi) - \langle \nabla f^*(\phi_t), \phi - \phi_t\rangle : \phi \in S \cap \ker(A^\top)\right\} \qquad (4.3)\\ &\leq \left(\frac{\left\| \left\| \phi_t - \mathsf{P}^{\|\|\cdot\|\|_*}_{S\cap\ker(A^\top)}(\phi_t) \right\|\right\|_*^2}{C_d L_t} \right)^{1/k}, \end{aligned}$$

where the last step used (C). Inserting this into (4.2) and subtracting $\inf_\lambda f(A\lambda)$,

$$f(A(\lambda_t + \alpha_t v_t)) - \inf_\lambda f(A\lambda) \leq f(A\lambda_t) - \inf_\lambda f(A\lambda) - \frac{C_d \gamma^2 (f(A\lambda_t) - \inf_\lambda f(A\lambda))^k}{6}. \tag{4.4}$$

When $k = 1$, recursively applying this expression gives a geometric sequence, and thus a rate $\mathcal{O}(\ln(1/\epsilon))$. When $k = 2$, a standard technique in optimization gives $\mathcal{O}(1/\epsilon)$ [9, Lemma 20]. □

Proof of Corollary 2.5. Adapting the proof of Proposition 13 from [1], it follows that $f + \iota_{\mathrm{im}(A)}$ is 0-coercive, meaning the initial level set $B := \{x \in \mathrm{im}(A) : f(x) \leq f(A\lambda_0)\}$ is compact. Since ∇f is continuously differentiable, $\nabla f(B)$ is compact, and strict convexity of f grants that $\nabla f(B) \subseteq \mathrm{int}(\mathrm{dom}(f^*))$; moreover it contains every dual iterate by construction, and the dual optimum by optimality conditions. As such, there exists a (compact) polytope S satisfying $\nabla f(B) \subseteq S \subseteq \mathrm{int}(\mathrm{dom}(f^*))$. The reverse map $\nabla f^*(S)$ is still compact (f^* is continuously differentiable by strict convexity [8, E.4.1.1]), and strict convexity of f means strong convexity over the compact set $\nabla f^*(S)$, thus (2.3) grants (C) with $k = 1$, and Theorem 2.4 gives the result. □

References

[1] Matus Telgarsky. A primal-dual convergence analysis of boosting. 2011. arXiv:1101.4752v2 [cs.LG].

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

[3] Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

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

[5] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. 2008. Submitted to SIAM J. Optim.

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

[7] 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.

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

[9] 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.

[10] Shai Shalev-Shwartz, Nathan Srebro, and Tong Zhang. Trading accuracy for sparsity. *SIAM Journal on Optimization*, 20(6):2807–2832, 2010.

[11] Alekh Agarwal, Sahand N. Negahban, and Martin J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. 2011. arXiv:1104.4824v1 [stat.ML].

[12] Jorge Nocedal and Stephen J. Wright. *Numerical optimization*. Springer, 2 edition, 2006.