Iterative Loss Minimization with ℓ_1-Norm Constraint and Guarantees on Sparsity

Shai Shalev-Shwartz and **Nathan Srebro**
Toyota Technological Institute—Chicago, USA
{shai,nati}@tti-c.org

July 3, 2008

Abstract

We study the problem of minimizing the loss of a linear predictor with a constraint on the ℓ_1 norm of the predictor. We describe a forward greedy selection algorithm for this task and analyze its rate of convergence. As a direct corollary of our convergence analysis we obtain a bound on the sparsity of the predictor as a function of the desired optimization accuracy, the bound on the ℓ_1 norm, and the Lipschitz constant of the loss function.

1 Outline of main results

We consider the problem of searching a linear predictor with low loss and low ℓ_1 norm. Formally, let $\mathcal{X}$ be an instance space, $\mathcal{Y}$ be a target space, and D be a distribution over $\mathcal{X} \times \mathcal{Y}$. Our goal is to approximately solve the following optimization problem

$$\min_{\mathbf{w}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[L(\langle \mathbf{w}, \mathbf{x}\rangle, y)] \quad \text{s.t.} \quad \|\mathbf{w}\|_1 \leq B \, , \tag{1}$$

where $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ is a loss function. Furthermore, we would like to find an approximated solution to Eq. (1) which is also sparse, namely, $\|\mathbf{w}\|_0 = |\{i : w_i \neq 0\}|$ is small.

We describe an iterative algorithm for solving Eq. (1) that alters a single element of $\mathbf{w}$ at each iteration. Assuming that L is convex and λ-Lipschitz with respect to its first argument, we prove that after performing T iterations of the algorithm it finds a solution with accuracy $O((\lambda\, B/\epsilon)^2)$. Our analysis therefore implies that we can find $\mathbf{w}$ such that

- $\|\mathbf{w}\|_0 = O((\lambda\, B/\epsilon)^2)$
- For all $\mathbf{w}^\star$ with $\|\mathbf{w}\|_1 \leq B$ we have $\mathbb{E}[L(\langle \mathbf{w}, \mathbf{x}\rangle, y)] \leq \mathbb{E}[L(\langle \mathbf{w}^\star, \mathbf{x}\rangle, y)] + \epsilon$

In a separate technical report, we show that this relation between $\|\mathbf{w}\|_0$, B, and ϵ is tight.

2 Problem Setting

Let $c : \mathbb{R}^n \to \mathbb{R}$ be the function

$$c(\mathbf{w}) \; = \; \mathbb{E}[L(\langle \mathbf{w}, \mathbf{x} \rangle, y)] \; .$$

Consider the problem

$$\min_{\mathbf{w}} c(\mathbf{w}) \quad \text{s.t.} \quad \|\mathbf{w}\|_1 \leq B \; , \tag{2}$$

and let $\mathbf{w}^\star$ be the minimizer of the above. Recall that our goal is to find a vector $\mathbf{w}$ such that $c(\mathbf{w}) - c(\mathbf{w}^\star) \leq \epsilon$ and $\|\mathbf{w}\|_0 = O(B^2/\epsilon^2)$.

In this report we present an iterative algorithm for solving Eq. (2). The algorithm initializes $\mathbf{w}_1 = \mathbf{0}$ and at each iteration it alters a single element of $\mathbf{w}$. Therefore, $\|\mathbf{w}_{t+1}\|_0 \leq \|\mathbf{w}_t\|_0 + 1$. We prove that the algorithm finds an ϵ-accurate solution of Eq. (2) after performing at most $O(B^2/\epsilon^2)$ iterations. As an immediate corollary we obtain that if we stop the procedure after performing $T = \Theta(B^2/\epsilon^2)$ iterations we will have $c(\mathbf{w}_T) \leq c(\mathbf{w}^\star) + \epsilon$ and $\|\mathbf{w}_T\|_0 \leq T$. That is, we obtain a sparsification procedure that finds a good sparse predictor without first finding a good low ℓ_1-norm predictor. Naturally, this procedure must be aware of the function c, that is, it should know (at least approximately) the distribution D and the loss function L. This stands in contrast to the randomized sparsification procedure described in the previous section, which is oblivious to D and L. Furthermore, to simplify our derivation we assume throughout this section that $\mathcal{D}$ is a distribution over a finite training set. Additionally, we assume that L is a proper convex function w.r.t. its first argument.

The report is organized as follows. Initially, we describe and analyze a forward greedy selection algorithm assuming that L has β Lipschitz continuous derivative (see Definition 1 below). We prove that the procedure finds an ϵ-accurate solution after performing at most $O(\frac{B^2}{\beta\,\epsilon})$ iterations. Next, we provide a mechanism for approximating any λ-Lipschitz function, L, by a function with β Lipschitz continuous derivative, $\tilde{L}$, with $\beta = \frac{\epsilon}{\lambda^2}$. This implies that we can run the forward greedy selection algorithm and find an $\epsilon/2$-accurate solution of $\tilde{c} = \mathbb{E}[\tilde{L}(\langle \mathbf{w}, \mathbf{x} \rangle, y)]$ after $O(\frac{\lambda^2\,B^2}{\epsilon^2})$ iterations. Combining this with the fact that $\tilde{c}$ approximates c, namely for all $\mathbf{w}$ $|c(\mathbf{w}) - \tilde{c}(\mathbf{w})| \leq \epsilon/2$, we obtain a guaranteed sparsification procedure for any λ-Lipschitz convex function.

Definition 1 *A loss function L has β Lipschitz continuous derivative if it is differentiable (w.r.t. its first argument) and its derivative (w.r.t. its first argument) satisfies*

$$\forall y \in \mathcal{Y}, \;\; \forall a_1, a_2 \in \mathbb{R}, \;\; |L'(a_1, y) - L'(a_2, y)| \leq \beta\, |a_1 - a_2| \; .$$

3 A forward greedy selection algorithm

We now describe a greedy forward selection algorithm for solving Eq. (2). The algorithm initializes the predictor vector to be the zero vector, $\mathbf{w}_1 = \mathbf{0}$. On iteration t, we first choose a feature by calculating the gradient of c at $\mathbf{w}_t$ (denoted $\boldsymbol{\theta}_t$) and finding

INPUT: Loss function $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$; ℓ_1 constraint B ;
Training set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ with $\|\mathbf{x}_i\|_\infty \le 1$ for all i

ASSUMPTION: L has β Lipschitz continuous derivative (see Definition 1)
(if not, see Sec. 4)

INITIALIZE: $\mathbf{w}_1 = \mathbf{0}$

FOR $t = 1, 2, \ldots$

$\boldsymbol{\theta}_t = \nabla c(\mathbf{w}_t)$ where $c(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^m L(\langle \mathbf{w}, \mathbf{x}_i\rangle, y_i)$

$j_t \in \arg\max_j |\theta_j|$
(w.l.o.g. assume $\mathrm{sign}(\theta)_{j_t} = -1$)

$\eta_t = \min\left\{1, \frac{\beta\,(\langle \boldsymbol{\theta}_t, \mathbf{w}_t\rangle + B\,\|\boldsymbol{\theta}_t\|_\infty)}{4\,B^2}\right\}$

$\mathbf{w}_{t+1} = (1-\eta_t)\mathbf{w}_t + \eta_t\, B\, \mathbf{e}^{j_t}$

STOPPING CONDITION: $\langle \boldsymbol{\theta}_t, \mathbf{w}_t\rangle + B\,\|\boldsymbol{\theta}_t\|_\infty \le \epsilon$

Figure 1: A greedy algorithm for solving Eq. (2) when L has β Lipschitz continuous derivative.

its largest element in absolute value. Then, we calculate a step size η_t and update the predictor according to

$$\mathbf{w}_{t+1} = (1-\eta_t)\mathbf{w}_t + \eta_t\, B\, \mathbf{e}^{j_t} \; .$$

The step size and the stopping criterion are based on our analysis below. Note that the update form ensures us that $\|\mathbf{w}_t\|_1 \le B$ and that $\|\mathbf{w}_t\|_0 \le t$. A pseudo-code describing the algorithm is given in Fig. 1.

The following theorem bounds the number of iterations required by the algorithm to converge.

Theorem 1 *Assume that the algorithm in Fig. 1 is run with a loss function L that has β Lipschitz continuous derivative and with a training set such that for all i, $\|\mathbf{x}_i\|_\infty \le 1$. Then, the algorithm stops after at most $O\left(\frac{B^2}{\beta\,\epsilon}\right)$ iterations.*

We now turn to the proof of Thm. 1. For all t, let ϵ_t be the sub-optimality of the algorithm at iteration t, that is,

$$\epsilon_t = c(\mathbf{w}_t) - \min_{\mathbf{w}:\|\mathbf{w}\|_1 \le B} c(\mathbf{w}) \; .$$

We also use $\mathbf{w}^\star$ to denote an optimal solution of Eq. (2).

The following lemma provides us with an upper bound on ϵ_t. Its proof using duality arguments (see the appendix for more details).

Lemma 1 *For all t we have $\langle \boldsymbol{\theta}_t, \mathbf{w}_t\rangle + B\,\|\boldsymbol{\theta}_t\|_\infty \;\ge\; \epsilon_t$.*

Proof From Fenchel duality, for any $\boldsymbol{\theta}$ we have

$$-c^\star(\boldsymbol{\theta}) - B\,\|\boldsymbol{\theta}\|_\infty \;\le\; \min_{\mathbf{w}:\|\mathbf{w}\|\le B} c(\mathbf{w}) \;\le\; c(\mathbf{w}_t)\;.$$

Therefore,

$$\epsilon_t \;\le\; c(\mathbf{w}_t) + c^\star(\boldsymbol{\theta}) + B\,\|\boldsymbol{\theta}\|_\infty$$

In particular, it holds for $\boldsymbol{\theta}_t = \nabla c(\mathbf{w}_t)$. But, in this case we also know from Lemma 7 that $c(\mathbf{w}_t) + c^\star(\boldsymbol{\theta}_t) = \langle \mathbf{w}_t, \boldsymbol{\theta}_t\rangle$. This concludes our proof. □

The next central lemma analyzes the progress of the algorithm.

Lemma 2 *Assume that L has β Lipschitz continuous derivative and that for all i, $\|\mathbf{x}_i\|_\infty \le 1$. Then,*

$$\epsilon_t - \epsilon_{t+1} \ge \eta_t\,\epsilon_t - \frac{2\,\eta_t^2\,B^2}{\beta}\;\;.$$

Proof Denote $\mathbf{u}_t = \eta_t(B\mathbf{e}^{j_t} - \mathbf{w}_t)$ and thus we can rewrite the update rule as $\mathbf{w}_{t+1} = (1-\eta_t)\mathbf{w}_t + \eta\,B\,\mathbf{e}^j = \mathbf{w}_t + \mathbf{u}_t$. Let $\Delta_t = \epsilon_t - \epsilon_{t+1} = c(\mathbf{w}_t) - c(\mathbf{w}_{t+1})$. Since L has β Lipschitz continuous derivative we can use Lemma 8 to get that for any $a_1, a_2 \in \mathbb{R}$ and $y \in \mathcal{Y}$ we have

$$L(a_1 + a_2, y) - L(a_1, y) \le L'(a_1)\,a_2 + \frac{a_2^2}{2\,\beta}\;. \tag{3}$$

Therefore,

$$\begin{aligned}
\Delta_t &= \frac{1}{m}\left(\sum_{i=1}^m \left(L(\langle \mathbf{w}_t, \mathbf{x}_i\rangle, y_i) - L(\langle \mathbf{w}_t + \mathbf{u}_t, x_i\rangle, y_i)\right)\right)\\
&\ge \frac{1}{m}\left(\sum_{i=1}^m \left(-L'(\langle \mathbf{w}_t, \mathbf{x}_i\rangle, y_i)\,\langle \mathbf{u}_t, x_i\rangle - \frac{(\langle \mathbf{u}_t, x_i\rangle)^2}{2\,\beta}\right)\right)\\
&= -\langle \boldsymbol{\theta}_t, \mathbf{u}_t\rangle - \frac{1}{m}\sum_{i=1}^m \frac{(\langle \mathbf{u}_t, x_i\rangle)^2}{2\,\beta}\;\;,
\end{aligned}$$

where the first equality follows from the definition of c, the second inequality follows from Eq. (3), and the in the last equality we used the definition of $\boldsymbol{\theta}_t$. Next, we use Holder inequality, the assumption $\|\mathbf{x}_i\|_\infty \le 1$, and the triangle inequality, to get that

$$\langle \mathbf{u}_t, \mathbf{x}_i\rangle \le \|\mathbf{u}_t\|_1 \|\mathbf{x}_i\|_\infty \le \|\mathbf{u}_t\|_1 \le \eta_t(\|B\mathbf{e}^{j_t}\|_1 + \|\mathbf{w}_t\|_1) \le 2\,\eta_t\,B\;\;.$$

Therefore,

$$\Delta_t \ge -\langle \boldsymbol{\theta}_t, \mathbf{u}_t\rangle - \frac{2\,\eta_t^2\,B^2}{\beta} \;=\; \eta_t\left(\langle \boldsymbol{\theta}_t, \mathbf{w}_t\rangle - B\,\langle \boldsymbol{\theta}_t, \mathbf{e}^{j_t}\rangle\right) - \frac{2\,\eta_t^2\,B^2}{\beta}\;\;. \tag{4}$$

The definition of j_t implies that $\langle \boldsymbol{\theta}_t, \mathbf{e}^{j_t}\rangle = -\|\boldsymbol{\theta}_t\|_\infty$. Therefore, we can invoke Lemma 1 and this concludes our proof. □

Equipped with the above lemma we are now ready to prove Thm. 1.

INPUT: Loss function $L : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$; ℓ_1 constraint B ; accuracy ϵ
ASSUMPTION: L is proper, convex, and λ-Lipschitz w.r.t. its first argument
STEP 1:

Set $\beta = \frac{\epsilon}{2\,\lambda^2}$

For each y define $\tilde{L}(\alpha, y) \;=\; \inf_v \; \frac{1}{2\,\beta} v^2 + L(\alpha - v, y)$

STEP 2:

Run the algorithm in Fig. 1 with $\tilde{L}$ and with accuracy $\frac{\epsilon}{2}$

Figure 2: A greedy algorithm for solving Eq. (2) for L being convex and λ-Lipschitz.

Proof [of Thm. 1] The definition of η_t implies that (see the proof of Lemma 2)

$$\Delta_t = \epsilon_t - \epsilon_{t+1} \geq \max_\eta \left(\eta \epsilon_t - \frac{2\eta^2 B^2}{\beta} \right) .$$

Note also that ϵ_t is monotonically decreasing. We consider two phases. At phase 1, we have $\epsilon_t > \frac{4\,B^2}{\beta}$. In this case, $\frac{\beta\,\epsilon_t}{4\,B^2} > 1$ and thus by setting $\eta = 1$ we obtain $\Delta_t \geq \frac{2\,B^2}{\beta}$. Therefore, the number of iterations in phase 1 is at most $\frac{\epsilon_1\,\beta}{2\,B^2} = O(1)$. At phase 2, we have $\epsilon_t \leq \frac{4\,B^2}{\beta}$ we can set $\eta = \frac{\beta\,\epsilon_t}{4\,B^2}$ and get that $\Delta_t \geq \frac{\beta\,\epsilon_t^2}{8\,B^2}$. Finally, Lemma 9 tells us that the number of iterations in phase 2 is at most $1 + \frac{8B^2}{\beta\,\epsilon}$. □

4 Approximating a Lipschitz-convex function by a function with a Lipschitz continuous gradient

Let $L : \mathbb{R} \to \mathbb{R}$ be a proper, convex, λ-Lipschitz function. The infimal convolution of L and the function $f(\alpha) = \frac{1}{2\,\beta}\|\alpha\|^2$ is defined as

$$\tilde{L}(\alpha) \;=\; \inf_v \; \frac{1}{2\,\beta} v^2 + L(\alpha - v) \,. \tag{5}$$

The following lemma states that $\tilde{L}$ approximates L and it has Lipschitz continuous gradient. Its proof is also useful for deriving a closed form of $\tilde{L}$ using the Fenchel conjugate operator.

Lemma 3 *Let L be a proper, convex, λ-Lipschitz function and let $\tilde{L}$ be as defined in Eq. (5). Then,*

- $\forall \alpha, \;\; |L(\alpha) - \tilde{L}(\alpha)| \leq \frac{\beta\,\lambda^2}{2}$
- *$\tilde{L}$ has β Lipschitz continuous gradient*

Proof Throughout the proof we use some definitions from convex analysis. In particular, the Fenchel conjugate of a function g is denoted by $g^\star$. See the appendix for more details. First, using Lemma 4 and the definition of the function f we know that

$$\tilde{L}^\star(\boldsymbol{\theta}) = f^\star(\theta) + L^\star(\theta) \;=\; \frac{\beta}{2}\theta^2 + L^\star(\theta)\ .$$

Therefore, $\tilde{L}^\star$ is β strongly convex (see appendix) and therefore using Lemma 8 we get that $\tilde{L}$ has β Lipschitz continuous gradient. This establishes the second claim of the lemma. Next, using Lemma 5 and the fact that L is λ Lipschitz we get that $\mathrm{dom}(L^\star) \subseteq [-\lambda, \lambda]$. Thus,

$$\tilde{L}^\star(\theta) \geq L^\star(\theta) = \tilde{L}^\star(\theta) - \frac{\beta\,\theta^2}{2} \;\geq\; \tilde{L}^\star(\theta) - \frac{\beta\,\lambda^2}{2}\ .$$

Finally, using Lemma 6 we conclude that

$$\tilde{L}(\alpha) \leq L(\alpha) \leq \tilde{L}(\alpha) + \frac{\beta\,\lambda^2}{2}\ .$$

$\square$

Based on the above lemma we obtain a following sparsification procedure that is applicable for any proper, convex, and λ Lipschitz loss function L. The sparsification procedure is outlined in Fig. 2. Combining Thm. 1 with Lemma 3 we obtain the following theorem:

Theorem 2 *If the sparsification procedure given in Fig. 2 is run with a proper, convex, and λ Lipschitz function L, then it finds $\mathbf{w}$ s.t. $c(\mathbf{w}) \leq c(\mathbf{w}^\star) + \epsilon$ and*

$$\|\mathbf{w}\|_0 = O\left(\frac{\lambda^2\,B^2}{\epsilon^2}\right)$$

Proof Using Thm. 1 and the definition of β we get that the output of the sparsification procedure satisfies

$$\|\mathbf{w}\|_0 \leq O\left(\frac{B^2}{\beta\,\epsilon}\right) \;=\; O\left(\frac{\lambda^2\,B^2}{\epsilon^2}\right)\ .$$

Let $\tilde{c}(\mathbf{w}) = \mathbb{E}[\tilde{L}(\langle\mathbf{w},\mathbf{x}\rangle, y)]$. Using Lemma 3, for any $\mathbf{w}$ we have

$$\begin{aligned}|\tilde{c}(\mathbf{w}) - c(\mathbf{w})| &= \left|\mathbb{E}[\tilde{L}(\langle\mathbf{w},\mathbf{x}\rangle, y) - L(\langle\mathbf{w},\mathbf{x}\rangle, y)]\right| \\ &\leq \left|\mathbb{E}[|\tilde{L}(\langle\mathbf{w},\mathbf{x}\rangle, y) - L(\langle\mathbf{w},\mathbf{x}\rangle, y)|\,]\right| \;\leq\; \frac{\beta\,\lambda^2}{2} = \frac{\epsilon}{4}\ .\end{aligned}$$

Let $\mathbf{w}^\star$ be the minimizer of $c(\mathbf{w})$ and let $\tilde{\mathbf{w}}^\star$ be the minimizer of $\tilde{c}(\mathbf{w})$. Then,

$$\begin{aligned}c(\mathbf{w}) - c(\mathbf{w}^\star) &= c(\mathbf{w}) - \tilde{c}(\mathbf{w}) + \tilde{c}(\mathbf{w}) - \tilde{c}(\mathbf{w}^\star) + \tilde{c}(\mathbf{w}^\star) - c(\mathbf{w}^\star) \\ &\leq \frac{\epsilon}{4} + \frac{\epsilon}{2} + \frac{\epsilon}{4} = \epsilon\ .\end{aligned}$$

This concludes our proof. $\square$

References

[BL06] J. Borwein and A. Lewis. *Convex Analysis and Nonlinear Optimization*. Springer, 2006.

[SS07] S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University, 2007.

[SSS06] S. Shalev-Shwartz and Y. Singer. On the equivalence of weak learnability and linear separability: New relaxations and efficient boosting algorithms. In *Proceedings of the Nineteenth Annual Conference on Computational Learning Theory*, 2006.

A Convex Analysis and Technical Lemmas

We first give a few basic definitions from convex analysis. We allow functions to output $+\infty$ and denote by $\mathrm{dom}(f)$ the set $\{\mathbf{w} : f(\mathbf{w}) < +\infty\}$. The Fenchel conjugate of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as

$$f^\star(\boldsymbol{\theta}) = \max_{\mathbf{w}} \ \langle \mathbf{w}, \boldsymbol{\theta} \rangle - f(\mathbf{w}) \ . \tag{6}$$

If f is closed and convex then $f^{\star\star} = f$.

The Fenchel weak duality theorem (see e.g. theorem 3.3.5 in [BL06]) states that for any two functions f, g we have

$$\max_{\boldsymbol{\theta}} -f^\star(-\boldsymbol{\theta}) - g^\star(\boldsymbol{\theta}) \ \leq \ \min_{\mathbf{w}} f(\mathbf{w}) + g(\mathbf{w}) \ .$$

The following lemma is a convolution theorem for infimal convolution.

Lemma 4 *If $f(\mathbf{w})$ and $g(\mathbf{w})$ are proper and convex functions and $h(\mathbf{w}) = \inf_{\mathbf{v}} f(\mathbf{v}) + g(\mathbf{w} - \mathbf{v})$ is their infimal convolution, then $h^\star = f^\star + g^\star$.*

The following lemma relates the Lipschitz property of c to the domain of its conjugate function.

Lemma 5 *If $c : \mathbb{R} \to \mathbb{R}$ is λ-Lipschitz then:* $\mathrm{dom}(c^\star) \subseteq [-\lambda, \lambda]$.

Proof From Lipschitz property we have $c(v) - c(0) \leq \lambda|v - 0| = \lambda|v|$ and thus $-c(v) \geq -(\lambda|v| + c(0))$. Therefore,

$$\begin{aligned} c^\star(\theta) &= \max_v \langle v, \theta \rangle - c(v) \\ &\geq \max_v \langle v, \theta \rangle - \lambda|v| - c(\mathbf{0}) = \begin{cases} \infty & \text{if } |\theta| > \lambda \\ -c(\mathbf{0}) & \text{else} \end{cases} \end{aligned}$$

□

Our next lemma is a perturbation lemma for Fenchel conjugate. Its proof can be found in [SSS06].

Lemma 6 *Let f, g be two functions and assume that for all $\mathbf{w} \in S$ we have $g(\mathbf{w}) \geq f(\mathbf{w}) \geq g(\mathbf{w}) - z$ for some constant z. Then, $g^\star(\boldsymbol{\theta}) \leq f^\star(\boldsymbol{\theta}) \leq g^\star(\boldsymbol{\theta}) + z$.*

The next lemma states a sufficient condition under which the Fenchel-Young inequality holds with equality. Its proof can be found in ([BL06], Proposition 3.3.4).

Lemma 7 *Let f be a closed and convex function and let $\partial f(\mathbf{w})$ be its differential set at $\mathbf{w}$. Then, for all $\boldsymbol{\theta} \in \partial f(\mathbf{w})$ we have, $f(\mathbf{w}) + f^\star(\boldsymbol{\theta}) = \langle \boldsymbol{\theta}, \mathbf{w} \rangle$.*

Next, we define the notion of strong convexity.

Definition 2 *A continuous function f is σ-strongly convex over a convex set S if S is contained in the domain of f and for all $\mathbf{v}, \mathbf{u} \in S$ and $\alpha \in [0, 1]$ we have*

$$\begin{aligned} f(\alpha\,\mathbf{v} + (1-\alpha)\,\mathbf{u}) \quad &\leq \quad \alpha\, f(\mathbf{v}) + (1-\alpha)\, f(\mathbf{u}) \\ &\quad -\frac{\sigma}{2}\,\alpha\,(1-\alpha)\,\|\mathbf{v}-\mathbf{u}\|^2 \ . \end{aligned}$$

The next lemma underscores the importance of strongly convex functions. For a proof see for example Lemma 18 in [SS07].

Lemma 8 *Let f be a proper and σ-strongly convex function over S. Let $f^\star$ be the Fenchel conjugate of f. Then, $f^\star$ has a σ Lipschitz continuous gradient. Furthermore, for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^n$, we have*

$$f^\star(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2) - f^\star(\boldsymbol{\theta}_1) \ \leq \ \langle \nabla f^\star(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2 \rangle + \frac{1}{2\sigma}\|\boldsymbol{\theta}_2\|^2$$

This technical lemma is used for proving the convergence of our greedy forward selection algorithm.

Lemma 9 *Let $r \in (0, 1/2)$ and let $\frac{1}{2r} \geq \epsilon_1 \geq \epsilon_2 \geq \ldots$ be a sequence such that for all $t \geq 1$ we have $\epsilon_t - \epsilon_{t+1} \geq r\,\epsilon_t^2$. Then, for all t we have $\epsilon_t \leq \frac{1}{r(t+1)}$.*

Proof We prove the lemma by induction. First, for $t = 1$ we have $\frac{1}{r(t+1)} = \frac{1}{2r}$ and the claim clearly holds. Assume that the claim holds for some t. Then,

$$\epsilon_{t+1} \leq \epsilon_t - r\epsilon_t^2 \leq \tfrac{1}{r(t+1)} - \tfrac{1}{r(t+1)^2} \quad , \tag{7}$$

where we used the fact that the function $x - rx^2$ is monotonically increasing in $[0, 1/(2r)]$ along with the inductive assumption. We can rewrite the right-hand side of Eq. (7) as

$$\tfrac{1}{r(t+2)}\left(\tfrac{(t+1)+1}{t+1} \cdot \tfrac{(t+1)-1}{t+1}\right) \ = \ \tfrac{1}{r(t+2)}\left(\tfrac{(t+1)^2-1}{(t+1)^2}\right) \ .$$

The term $\frac{(t+1)^2-1}{(t+1)^2}$ is smaller than 1 and thus $\epsilon_{t+1} \leq \frac{1}{r(t+2)}$, which concludes our proof.