Logarithmic Regret Algorithms for *Strongly* Convex Repeated Games

Shai Shalev-Shwartz[1] and Yoram Singer[1,2]
[1] School of Computer Sci. & Eng., The Hebrew University, Jerusalem 91904, Israel
[2] Google Inc. 1600 Amphitheater Parkway, Mountain View, CA 94043, USA

May 20, 2007

Abstract

Many problems arising in machine learning can be cast as a convex optimization problem, in which a sum of a loss term and a regularization term is minimized. For example, in Support Vector Machines the loss term is the average hinge-loss of a vector over a training set of examples and the regularization term is the squared Euclidean norm of this vector. In this paper we study an algorithmic framework for strongly convex repeated games and apply it for solving regularized loss minimization problems. In a convex repeated game, a predictor chooses a sequence of vectors from a convex set. After each vector is chosen, the opponent responds with a convex loss function and the predictor pays for applying the loss function to the vector she chose. The regret of the predictor is the difference between her cumulative loss and the minimal cumulative loss achievable by a fixed vector, even one that is chosen in hindsight. In strongly convex repeated games, the opponent is forced to choose loss functions that are strongly convex. We describe a family of prediction algorithms for strongly convex repeated games that attain logarithmic regret.

1 Introduction

A convex repeated game is a two players game that is performed in a sequence of consecutive rounds. On round t of the repeated game, the first player chooses a vector $\mathbf{w}_t$ from a convex set S. Next, the opponent responds with a convex loss function $g_t : S \to \mathbb{R}$. Finally, the first player suffers an instantaneous loss $g_t(\mathbf{w}_t)$. We study the game from the viewpoint of the first player which we also call the predictor. The goal of the predictor is to minimize its regret. Formally, the regret after T iterations of the game is defined to be

$$\sum_{t=1}^{T} g_t(\mathbf{w}_t) - \min_{\mathbf{u} \in S} \sum_{t=1}^{T} g_t(\mathbf{u}).$$

In strongly convex repeated games, each function g_t must be σ-strongly convex. Intuitively, the scalar σ measures how much the function g_t stretches above its tangents. The formal definition is given in the next section. For strongly convex repeated games, we propose a strategy for the predictor that guarantees a regret bound of $O(\frac{\log(T)\, L}{\sigma})$, where L is a (generalized) Lipschitz coefficient of the functions $\{g_t\}$ that will be defined in later sections.

Strongly convex repeated games are instrumental for solving regularized loss minimization problems of the form

$$\min_{\mathbf{w}\in S} \sigma f(\mathbf{w}) + \frac{1}{m}\sum_{i=1}^{m} \ell_i(\mathbf{w}) \,. \tag{1}$$

For example, in Support vector machines $f(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$ and $\ell_i(\mathbf{w}) = \max\{0, 1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle\}$ for some pair $(\mathbf{x}_i, y_i) \in \mathbb{R}^n \times \{+1, -1\}$. The parameter σ is a non-negative scalar that balances the tradeoff between the loss and regularization terms. Denote by $P(\mathbf{w})$ the objective function in Eq. (1). As we show in the next section, a function g that can be written as a sum $g(\mathbf{w}) = \sigma f(\mathbf{w}) + h(\mathbf{w})$, where $f(\cdot)$ and $g(\cdot)$ are convex, is σ-strongly convex w.r.t. $f(\cdot)$. Assuming that the loss functions $\{\ell_i\}$ are convex we obtain that $P(\mathbf{w})$ is σ-strongly convex w.r.t. $f(\cdot)$. Therefore, setting $g_t(\cdot) = P(\cdot)$ for all t and using a strategy for the predictor that attain logarithmic regret we immediately obtain a solver for Eq. (1) for which

$$\min_{t\in[T]} P(\mathbf{w}_t) \;\le\; \frac{1}{T}\sum_{t=1}^{T} P(\mathbf{w}_t) \;\le\; \min_{\mathbf{u}\in S} P(\mathbf{u}) + O\left(\frac{\log(T)\, L}{T\,\sigma}\right) \,.$$

The cost of each iteration of the above procedure scales linearly with m. An alternative approach is to set $g_t = \sigma f(\mathbf{w}) + \ell_i(\mathbf{w})$, where i is chosen randomly from $[m]$. For this approach, we are able to obtain a similar convergence rate that holds with high probability.

The rest of this paper is organized as follows. We start in Sec. 2 by establishing our notation and pointing to a few mathematical tools used throughout the paper. Our algorithmic framework for strongly convex repeated games is presented and analyzed in Sec. 3. Next, in Sec. 4, we outline the applicability of the framework for solving regularized loss minimization problems.

2 Mathematical Background

In this section we establish our notation and give references to a few mathematical tools used throughout the paper. We denote scalars with lower case letters (e.g. x), and vectors with bold face letters (e.g. $\mathbf{x}$). The inner product between vectors $\mathbf{x}$ and $\mathbf{w}$ is denoted by $\langle \mathbf{x}, \mathbf{w}\rangle$. Sets are designated by upper case letters (e.g. S). The set of non-negative real numbers is denoted by $\mathbb{R}_+$. For any $k \ge 1$, the set of integers $\{1, \ldots, k\}$ is denoted by $[k]$. A norm of a vector $\mathbf{x}$ is denoted by $\|\mathbf{x}\|$. The dual norm is defined as $\|\boldsymbol{\lambda}\|_\star = \sup\{\langle \mathbf{x}, \boldsymbol{\lambda}\rangle : \|\mathbf{x}\| \le 1\}$. For example, the Euclidean norm, $\|\mathbf{x}\|_2 = (\langle \mathbf{x}, \mathbf{x}\rangle)^{1/2}$ is dual to itself and the ℓ_1 norm, $\|\mathbf{x}\|_1 = \sum_i |x_i|$, is dual to the ℓ_∞ norm, $\|\mathbf{x}\|_\infty = \max_i |x_i|$.

We next recall a few definitions from convex analysis. The reader familiar with convex analysis may proceed to Definition 1. For a more thorough introduction see for example [1, 4]. A set S is convex if for any two vectors $\mathbf{w}_1, \mathbf{w}_2$ in S, all the line between $\mathbf{w}_1$ and $\mathbf{w}_2$ is also within S. That is, for any $\alpha \in [0, 1]$ we have that $\alpha\mathbf{w}_1 + (1 - \alpha)\mathbf{w}_2 \in S$. A set S is open if every point in S has a neighborhood lying in S. A set S is closed if its complement is an open set. A function $f : S \to \mathbb{R}$ is closed and convex if for any scalar $\alpha \in \mathbb{R}$, the level set $\{\mathbf{w} : f(\mathbf{w}) \le \alpha\}$ is closed and convex.

The Fenchel conjugate of a function $f : S \to \mathbb{R}$ is defined as

$$f^\star(\boldsymbol{\theta}) \;=\; \sup_{\mathbf{w}\in S} \; (\langle \mathbf{w}, \boldsymbol{\theta}\rangle - f(\mathbf{w})) \,.$$

Since $f^\star$ is defined as a supremum of linear functions it is convex. If in addition f is closed and convex then the Fenchel conjugate of $f^\star$ is f itself. Throughout this paper we work solely with functions that are close.

A vector $\boldsymbol{\lambda}$ is a sub-gradient of a function f at $\mathbf{v}$ if

$$\forall \mathbf{u} \in S, \;\; f(\mathbf{u}) - f(\mathbf{v}) \geq \langle \mathbf{u} - \mathbf{v}, \boldsymbol{\lambda} \rangle \; .$$

The differential set of f at $\mathbf{v}$, denoted $\partial f(\mathbf{v})$, is the set of all sub-gradients of f at $\mathbf{v}$. A function f is convex iff $\partial f(\mathbf{v})$ is non-empty for all $\mathbf{v} \in S$. If f is convex and differentiable at $\mathbf{v}$ then $\partial f(\mathbf{v})$ consists of a single vector which amounts to the gradient of f at $\mathbf{v}$ and is denoted by $\nabla f(\mathbf{v})$. As a consequence we obtain that a differential function f is convex iff for all $\mathbf{v}, \mathbf{u} \in S$ we have that

$$f(\mathbf{u}) - f(\mathbf{v}) - \langle \mathbf{u} - \mathbf{v}, \nabla f(\mathbf{v}) \rangle \geq 0 \; .$$

The left-hand side of the above inequality is called the Bregman divergence between $\mathbf{u}$ and $\mathbf{v}$ and is denoted as

$$B_f(\mathbf{u} \| \mathbf{v}) \;=\; f(\mathbf{u}) - f(\mathbf{v}) - \langle \mathbf{u} - \mathbf{v}, \nabla f(\mathbf{v}) \rangle \; . \tag{2}$$

For example, the function $f(\mathbf{v}) = \frac{1}{2}\|\mathbf{v}\|_2^2$ yields the divergence $B_f(\mathbf{u}\|\mathbf{v}) = \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|_2^2$. Since f is convex, the Bregman divergence is non-negative.

The focus of this paper is on strongly convex functions.

Definition 1 *A closed and convex function f is σ-***strongly** *convex over S with respect to a norm $\|\cdot\|$ if*

$$\forall \mathbf{u}, \mathbf{v} \in S, \; \forall \boldsymbol{\lambda} \in \partial f(\mathbf{v}), \;\; f(\mathbf{u}) - f(\mathbf{v}) - \langle \mathbf{u} - \mathbf{v}, \boldsymbol{\lambda} \rangle \;\geq\; \sigma \, \frac{\|\mathbf{u} - \mathbf{v}\|^2}{2} \; .$$

A direct generalization of the above definition is by replacing the squared norm at the right-hand side of the inequality with a Bregman divergence.

Definition 2 *A convex function g is σ-***strongly** *convex over S with respect to a convex and differentiable function f if*

$$\forall \mathbf{u}, \mathbf{v} \in S, \; \forall \boldsymbol{\lambda} \in \partial g(\mathbf{v}), \;\; g(\mathbf{u}) - g(\mathbf{v}) - \langle \mathbf{u} - \mathbf{v}, \boldsymbol{\lambda} \rangle \;\geq\; \sigma B_f(\mathbf{u} \| \mathbf{v}) \; .$$

The following lemma provides a sufficient condition for strong convexity of g w.r.t. f.

Lemma 1 *Assume that f is a differentiable and convex function and let $g = \sigma f + h$ where h is also a convex function. Then, g is σ-strongly convex w.r.t. f.*

Proof Let $\mathbf{v}, \mathbf{u} \in S$ and choose a vector $\boldsymbol{\lambda} \in \partial g(\mathbf{v})$. Since $\partial g(\mathbf{v}) = \partial h(\mathbf{v}) + \sigma \partial f(\mathbf{v})$, we have that there exists $\boldsymbol{\lambda}_1 \in \partial h(\mathbf{v})$ s.t. $\boldsymbol{\lambda} = \boldsymbol{\lambda}_1 + \nabla f(\mathbf{v})$. Thus,

$$g(\mathbf{u}) - g(\mathbf{v}) - \langle \mathbf{u} - \mathbf{v}, \boldsymbol{\lambda} \rangle = B_f(\mathbf{u} \| \mathbf{v}) + h(\mathbf{u}) - h(\mathbf{v}) - \langle \boldsymbol{\lambda}_1, \mathbf{u} - \mathbf{v} \rangle \geq B_f(\mathbf{u} \| \mathbf{v}) \; ,$$

where the last inequality follows from the convexity of h. ■

When a function is 1-strongly convex, we often omit the reference to the constant 1. Two notable examples of strongly convex functions which we use are given in the following examples.

Example 1 *The function* $f(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$ *is strongly convex over* $S = \mathbb{R}^n$ *with respect to the* ℓ_2 *norm. Its conjugate function is* $f^\star(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|_2^2$*. More generally, for* $q > 1$*, the function* $f(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_q^2$ *is strongly convex over* $S = \mathbb{R}^n$ *with respect to the* ℓ_q *norm. Its conjugate function is* $f^\star(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{\theta}\|_p^2$*, where* $\frac{1}{p} + \frac{1}{q} = 1$*.*

Example 2 *The function* $f(\mathbf{w}) = \sum_{i=1}^n w_i \log(\frac{w_i}{1/n})$ *is strongly convex over the probabilistic simplex,* $S = \{\mathbf{w} \in \mathbb{R}_+^n : \|\mathbf{w}\|_1 = 1\}$*, with respect to the* ℓ_1 *norm. For a proof see Lemma 8 in Appendix A. Its conjugate function is* $f^\star(\boldsymbol{\theta}) = \log(\frac{1}{n}\sum_{i=1}^n \exp(\theta_i))$*.*

Strongly convex functions play an important role in our analysis mainly due to the following lemma.

Lemma 2 *Let* f *be a differential and* σ*-strongly convex function over* S *with respect to a norm* $\|\cdot\|$ *. Then,*

1. $f^\star$ *is differentiable and* $\nabla f^\star(\boldsymbol{\theta}) = \operatorname*{argmax}_{\mathbf{w}\in S} \langle \mathbf{w}, \boldsymbol{\theta}\rangle - f(\mathbf{w})$ *.*
2. $\forall \boldsymbol{\theta} \in \mathbb{R}^n,\ \forall \mathbf{u} \in S,\ \ \langle \mathbf{u} - \nabla f^\star(\boldsymbol{\theta}), \boldsymbol{\theta} - \nabla f(\nabla f^\star(\boldsymbol{\theta}))\rangle \le 0$ *.*

Proof The first claim is Lemma 6 in Appendix A. Denote $\mathbf{v} = \nabla f^\star(\boldsymbol{\theta})$, we have that

$$\mathbf{v} = \operatorname*{argmax}_{\mathbf{w}\in S} \langle \mathbf{w}, \boldsymbol{\theta}\rangle - f(\mathbf{w}) \,.$$

Denote the objective of the maximization problem by $P(\mathbf{w})$. The assumption that f is differentiable implies that P is also differentiable with $\nabla P(\mathbf{w}) = \boldsymbol{\theta} - \nabla f(\mathbf{w})$. The optimality of $\mathbf{v}$ implies that for all $\mathbf{u} \in S$

$$\langle \mathbf{u} - \mathbf{v}, \nabla P(\mathbf{v})\rangle \le 0 \,,$$

which concludes our proof since $\nabla P(\mathbf{v}) = \boldsymbol{\theta} - \nabla f(\mathbf{v})$. ■

3 A Logarithmic Regret Algorithmic Framework for Strongly Convex Repeated Games

In this section we describe our algorithmic framework for playing strongly convex repeated games. Recall that we study the game from the viewpoint of the predictor and would like to have a logarithmic regret bound. The predictor constructs her sequence of vectors as follows:

- Parameters: A function $f : S \to \mathbb{R}$ and a scalar $\sigma > 0$
- Initialize: $\mathbf{w}_1 \in S$
- For $t = 1, 2, \ldots, T$
  - Play: $\mathbf{w}_t$
  - Receive function g_t from opponent
  - Update:

1. Choose $\boldsymbol{\lambda}_t \in \partial g_t(\mathbf{w}_t)$
2. Set $\eta_t = 1/(\sigma t)$
3. Set $\mathbf{w}_{t+1} = \nabla f^\star\left(\nabla f(\mathbf{w}_t) - \eta_t \boldsymbol{\lambda}_t\right)$

Before we turn to the analysis of the algorithm, let us first describe a couple of specific examples which we also use in Sec. 4.

Example 3 *Let S be a closed convex set and let $f(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$. The conjugate of f is,*

$$f^\star(\boldsymbol{\theta}) \;=\; \max_{\mathbf{w}\in S} \; \langle \mathbf{w}, \boldsymbol{\theta}\rangle - \frac{1}{2}\|\mathbf{w}\|_2^2 \;=\; \frac{1}{2}\|\boldsymbol{\theta}\|_2^2 - \min_{\mathbf{w}\in S} \frac{1}{2}\|\mathbf{w}-\boldsymbol{\theta}\|_2^2 \,.$$

Based on Lemma 2 we also obtain that $\nabla f^\star(\boldsymbol{\theta})$ is the projection of $\boldsymbol{\theta}$ onto S, that is,

$$\nabla f^\star(\boldsymbol{\theta}) \;=\; \operatorname*{argmin}_{\mathbf{w}\in S} \|\mathbf{w}-\boldsymbol{\theta}\|_2^2 \,.$$

Therefore, the update of our algorithm can be written as

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}\in S} \|\mathbf{w} - (\mathbf{w}_t - \eta_t \boldsymbol{\lambda}_t)\|_2^2 \;\; .$$

When g_t is differentiable, this specific update procedure was previously proposed in [2]. Note that when S is the n'th dimensional ball of radius ρ, $S = \{\mathbf{w} \,|\, \|\mathbf{w}\| \leq \rho\}$, the projection of $\boldsymbol{\theta}$ on S amounts to scaling $\boldsymbol{\theta}$ by $\min\{1, \frac{\rho}{\|\boldsymbol{\theta}\|}\}$.

Example 4 *Let S be the n'th dimensional probability simplex,*

$$S = \{\mathbf{w} | \sum_{j=1}^n w_j = 1 \,,\, \forall j : \, w_j \geq 0\} \;\; ,$$

and let $f(\mathbf{w}) = \sum_{j=1}^n w_j \log(w_j) + \log(n)$. The conjugate of f is,

$$\begin{aligned} f^\star(\boldsymbol{\theta}) \;&=\; \max_{\mathbf{w}\in S} \; \langle \mathbf{w}, \boldsymbol{\theta}\rangle - \sum_{j=1}^n w_j \log(w_j) - \log(n) \\ &=\; -\log(n) - \min_{\mathbf{w}\in S} \sum_{j=1}^n w_j \log\left(\frac{w_j}{e^{\theta_j}}\right) \;\; . \end{aligned}$$

Using again Lemma 2 we obtain that $\nabla f^\star(\boldsymbol{\theta})$ is the (entropic) projection of $\boldsymbol{\theta}$ onto the simplex S, that is,

$$\nabla f^\star(\boldsymbol{\theta}) \;=\; \operatorname*{argmin}_{\mathbf{w}\in S} \sum_{j=1}^n w_j \log\left(\frac{w_j}{e^{\theta_j}}\right) \;\; .$$

Algebraic manipulations yield that step (3), namely $\mathbf{w}_{t+1} = \nabla f^\star\left(\nabla f(\mathbf{w}_t) - \eta_t \boldsymbol{\lambda}_t\right)$, when $\nabla f^\star(\boldsymbol{\theta})$ is of the above form, can be written as follows,

$$w_{t+1,j} = \frac{w_{t,j} e^{-\eta_t \lambda_{t,j}}}{Z_t} \;\; \textit{where} \;\; Z_t = \sum_{r=1}^n w_{t,r} e^{-\eta_t \lambda_{t,r}} \;\; .$$

In this case we recovered a well known version of the exponentiated gradient (EG) algorithm [3]. The set S can be further restricted. In Sec. 4 we describe an algorithm that uses the function f of this example with a more restricted domain S defined as

$$S = \{\mathbf{w} | \sum_{j=1}^{n} w_j = 1 \,,\, \forall j : \, w_j \geq \epsilon\} \ .$$

In App. B we describe an efficient algorithm for solving step (3) for this more complex domain S.

3.1 Analysis

In this section we analyze our algorithmic framework. Specifically, we show that if the opponent chooses σ-strongly convex functions w.r.t. f then the regret of the predictor is logarithmic. First, we need the following lemma.

Lemma 3 *Let f be a 1-strongly convex function w.r.t. a norm $\|\cdot\|$ over S and $\mathbf{u}$ be an arbitrary vector in S. Then,*

$$\langle \mathbf{w}_t - \mathbf{u}, \boldsymbol{\lambda}_t \rangle \leq \frac{B_f(\mathbf{u} \| \mathbf{w}_t) - B_f(\mathbf{u} \| \mathbf{w}_{t+1})}{\eta_t} + \eta_t \frac{\|\boldsymbol{\lambda}_t\|_\star^2}{2} \ .$$

Proof Denote $\Delta_t = B_f(\mathbf{u} \| \mathbf{w}_t) - B_f(\mathbf{u} \| \mathbf{w}_{t+1})$. Expanding the definition of B_f we have that

$$\Delta_t = \langle \mathbf{u} - \mathbf{w}_{t+1}, \nabla f(\mathbf{w}_{t+1}) - \nabla f(\mathbf{w}_t) \rangle + B_f(\mathbf{w}_{t+1} \| \mathbf{w}_t) \ .$$

Since f is 1-strongly convex w.r.t. $\|\cdot\|$ we have that

$$\Delta_t \geq \langle \mathbf{u} - \mathbf{w}_{t+1}, \nabla f(\mathbf{w}_{t+1}) - \nabla f(\mathbf{w}_t) \rangle + \frac{1}{2} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \ . \tag{3}$$

Let us denote by $\boldsymbol{\theta}_t$ the term $\nabla f(\mathbf{w}_t) - \eta_t \boldsymbol{\lambda}_t$. Using the second part of Lemma 2 with $\boldsymbol{\theta}$ set to be $\boldsymbol{\theta}_t$ and rewriting $\nabla f^\star(\boldsymbol{\theta}_t)$ as $\mathbf{w}_{t+1}$ (see step 3 above) we get that,

$$\begin{aligned} 0 &\geq \langle \mathbf{u} - \nabla f^\star(\boldsymbol{\theta}_t), \boldsymbol{\theta}_t - \nabla f(\nabla f^\star(\boldsymbol{\theta}_t)) \rangle \\ &= \langle \mathbf{u} - \mathbf{w}_{t+1}, \boldsymbol{\theta}_t - \nabla f(\mathbf{w}_{t+1}) \rangle \\ &= \langle \mathbf{u} - \mathbf{w}_{t+1}, \nabla f(\mathbf{w}_t) - \eta_t \boldsymbol{\lambda}_t - \nabla f(\mathbf{w}_{t+1}) \rangle \ . \end{aligned}$$

Rearranging terms we we obtain that

$$\langle \mathbf{u} - \mathbf{w}_{t+1}, \nabla f(\mathbf{w}_{t+1}) - \nabla f(\mathbf{w}_t) \rangle \ \geq \ \eta_t \langle \mathbf{w}_{t+1} - \mathbf{u}, \boldsymbol{\lambda}_t \rangle \ .$$

Combining the above with Eq. (3) gives that

$$\begin{aligned} \Delta_t \ &\geq \ \eta_t \langle \mathbf{w}_{t+1} - \mathbf{u}, \boldsymbol{\lambda}_t \rangle + \frac{1}{2} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \\ &= \ \eta_t \langle \mathbf{w}_t - \mathbf{u}, \boldsymbol{\lambda}_t \rangle - \langle \mathbf{w}_{t+1} - \mathbf{w}_t, \eta_t \boldsymbol{\lambda}_t \rangle + \frac{1}{2} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \ . \end{aligned}$$

Applying the inequality

$$|\langle \mathbf{v}, \boldsymbol{\theta} \rangle| \leq \|\mathbf{v}\| \, \|\boldsymbol{\theta}\|_\star = \frac{1}{2}\|\mathbf{v}\|^2 + \frac{1}{2}\|\boldsymbol{\theta}\|_\star^2 - \frac{1}{2}\left(\|\mathbf{v}\| - \|\boldsymbol{\theta}\|_\star\right)^2 \leq \frac{1}{2}\|\mathbf{v}\|^2 + \frac{1}{2}\|\boldsymbol{\theta}\|_\star^2 \,,$$

which holds for all $\mathbf{v}$ and $\boldsymbol{\theta}$, we obtain that

$$\begin{aligned}\Delta_t &\geq \eta_t \langle \mathbf{w}_t - \mathbf{u}, \boldsymbol{\lambda}_t \rangle - \frac{1}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 - \frac{1}{2}\|\eta_t \boldsymbol{\lambda}_t\|_\star^2 + \frac{1}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \\ &= \eta_t \langle \mathbf{w}_t - \mathbf{u}, \boldsymbol{\lambda}_t \rangle - \frac{\eta_t^2}{2}\|\boldsymbol{\lambda}_t\|_\star^2 \,.\end{aligned}$$

■

Theorem 1 *Let f be a 1-strongly convex function w.r.t. a norm $\|\cdot\|$ over S. Assume that for all t, g_t is a σ-strongly convex function w.r.t. f. Additionally, let L be a scalar such that $\frac{1}{2}\|\boldsymbol{\lambda}_t\|_\star^2 \leq L$ for all t. Then, the following bound holds for all $T \geq 1$,*

$$\sum_{t=1}^{T} g_t(\mathbf{w}_t) - \sum_{t=1}^{T} g_t(\mathbf{u}) \;\leq\; \frac{L}{\sigma}\left(1 + \log(T)\right) .$$

Proof From Lemma 3 we have that

$$\langle \mathbf{w}_t - \mathbf{u}, \boldsymbol{\lambda}_t \rangle \leq \frac{B_f(\mathbf{u}\|\mathbf{w}_t) - B_f(\mathbf{u}\|\mathbf{w}_{t+1})}{\eta_t} + \eta_t \frac{\|\boldsymbol{\lambda}_t\|_\star^2}{2} \,. \tag{4}$$

Since g_t is σ-strongly convex w.r.t. f we can bound the left-hand side of the above inequality as follows,

$$\langle \mathbf{w}_t - \mathbf{u}, \boldsymbol{\lambda}_t \rangle \geq g_t(\mathbf{w}_t) - g_t(\mathbf{u}) + \sigma B_f(\mathbf{u}\|\mathbf{w}_t) \,. \tag{5}$$

Combining Eq. (4) with Eq. (5), and using the assumption $\frac{1}{2}\|\boldsymbol{\lambda}_t\|_\star^2 \leq L$ we get that

$$g_t(\mathbf{w}_t) - g_t(\mathbf{u}) \;\leq\; \left(\frac{1}{\eta_t} - \sigma\right) B_f(\mathbf{u}\|\mathbf{w}_t) - \frac{1}{\eta_t} B_f(\mathbf{u}\|\mathbf{w}_{t+1}) + \eta_t L \,.$$

Summing over t we obtain

$$\begin{aligned}\sum_{t=1}^{T} \left(g_t(\mathbf{w}_t) - g_t(\mathbf{u})\right) \;\leq\; &\left(\frac{1}{\eta_1} - \sigma\right) B_f(\mathbf{u}\|\mathbf{w}_1) - \left(\frac{1}{\eta_T}\right) B_f(\mathbf{u}\|\mathbf{w}_{T+1}) + \\ &\sum_{t=2}^{T} B_f(\mathbf{u}\|\mathbf{w}_t) \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \sigma\right) + L \sum_{t=1}^{T} \eta_t \,.\end{aligned}$$

Plugging the value of η_t in the above inequality, we obtain that the first and third summands on the right-hand side of the inequality vanish and that the second summand is negative. We therefore get,

$$\sum_{t=1}^{T} \left(g_t(\mathbf{w}_t) - g_t(\mathbf{u})\right) \;\leq\; L \sum_{t=1}^{T} \eta_t \;=\; \frac{L}{\sigma} \sum_{t=1}^{T} \frac{1}{t} \;\leq\; \frac{L}{\sigma}(\log(T) + 1) \,.$$

■

4 Regularized Loss Minimization

In this section we describe an application of our algorithmic framework from the previous section for solving regularized loss minimization problems of the form given in Eq. (1). For concreteness, we describe the usage of our framework for finding a solution for a classification problem with the hinge-loss and with ℓ_2 regularization. This problem was widely studied and is known as the support vector machine (SVM) problem. The second setting we discuss is logistic regression for classification with ℓ_1 regularization. We would like to note though that our approach is clearly applicable to other learning problems such as regression learning with other regularization functions that adhere with our setting.

Our first concrete derivation focuses on solving the SVM optimization problem, which is defined as follows,

$$\min_{\mathbf{w}\in\mathbb{R}^n} \frac{\sigma}{2}\|\mathbf{w}\|_2^2 + \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle\} \;, \tag{6}$$

where for all $i \in [m]$ we have $(\mathbf{x}_i, y_i) \in \mathbb{R}^n \times \{-1, +1\}$. Denote by $g(\mathbf{w})$ the objective function in Eq. (6). Additionally, let $S = \{\mathbf{w} : \|\mathbf{w}\|_2 \leq 1/\sqrt{\sigma}\}$. The following lemma shows that the problem in Eq. (6) is equivalent to the problem.

$$\min_{\mathbf{w}\in S} \frac{\sigma}{2}\|\mathbf{w}\|_2^2 + \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle\} \;. \tag{7}$$

Note that the sole difference between Eq. (6) and Eq. (7) is the additional restriction of $\mathbf{w}$ to the domain $S \subset \mathbb{R}^n$.

Lemma 4 *The norm of the optimum of the optimization problem defined in Eq. (6) is bounded above by* $1/\sqrt{\sigma}$.

Proof Let us denote the optimal solution of Eq. (6) by $\mathbf{w}^\star$. The dual problem of the optimization problem defined in Eq. (6) is

$$\max_{\boldsymbol{\alpha}\in[0,1/m]^m} \sum_{i=1}^{m} \alpha_i - \frac{1}{2\sigma}\Big\|\sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i\Big\|^2 \;.$$

Let $\boldsymbol{\alpha}^\star$ be an optimal solution of the dual problem. Since strong duality holds, we obtain that at the optimum the dual objective and the primal objective coincide, that is,

$$\frac{\sigma}{2}\|\mathbf{w}^\star\|^2 + \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle\} = \sum_{i=1}^{m} \alpha_i^\star - \frac{1}{2\sigma}\Big\|\sum_{i=1}^{m} \alpha_i^\star y_i \mathbf{x}_i\Big\|^2 \;. \tag{8}$$

In addition, at the optimum we have that $\mathbf{w}^\star = \frac{1}{\sigma}\sum_{i=1}^{m} \alpha_i^\star y_i \mathbf{x}_i$. Plugging this equality into Eq. (8) and rearranging terms yields

$$\sigma\|\mathbf{w}^\star\|^2 = \sum_{i=1}^{m} \alpha_i^\star - \frac{1}{m}\sum_{i=1}^{m} \max\{0, 1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle\} \leq 1 \;.$$

■

We now turn to the description of the algorithm for solving Eq. (7) based on our framework from Sec. 3. Initially, we set $\mathbf{w}_1$ to any vector in S. On iteration t of the algorithm, we first choose a set $I_t \subseteq [m]$. Then, we replace the objective in Eq. (7) with an instantaneous objective function,

$$g_t(\mathbf{w}) \;=\; \frac{\sigma}{2}\|\mathbf{w}\|^2 + \frac{1}{|I_t|}\sum_{i\in I_t}\max\{0, 1 - y_i\langle \mathbf{w}, \mathbf{x}_i\rangle\} \;\;.$$

Note that g_t is σ-strongly convex w.r.t. $f(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$ (see Lemma 1). Next, we set the learning rate $\eta_t = 1/(\sigma t)$ and update

$$\mathbf{w}_{t+1} = \min_{\mathbf{w}\in S} \|\mathbf{w} - (\mathbf{w}_t - \eta_t \boldsymbol{\lambda}_t)\|_2 \;, \tag{9}$$

where $\boldsymbol{\lambda}_t \in \partial g_t(\mathbf{w}_t)$. For example, we can set

$$\boldsymbol{\lambda}_t \;=\; \sigma\,\mathbf{w}_t - \frac{1}{|I_t|}\sum_{i:y_i\langle \mathbf{w}_t,\mathbf{x}_i\rangle<1} y_i\,\mathbf{x}_i \;.$$

Assume that $\|\mathbf{x}_i\| \le R$ for all i, then we have

$$\|\boldsymbol{\lambda}_t\| \le \sigma\|\mathbf{w}_t\| + R \le \sqrt{\sigma} + R \;.$$

Based on Example 3 and Thm. 1 we obtain the following corollary.

Corollary 1 *Assume that for all $i \in [m]$ the norm of $\mathbf{x}_i$ is at most R. Let $\mathbf{w}_1, \ldots, \mathbf{w}_T$ be defined according to Eq. (9) and let $\mathbf{u}$ be an arbitrary vector in S. Then,*

$$\sum_{t=1}^{T} g_t(\mathbf{w}_t) \;\le\; \sum_{t=1}^{T} g_t(\mathbf{u}) + \frac{(\sqrt{\sigma}+R)^2}{2\sigma}\left(1+\log(T)\right) \;\;.$$

If we set $I_t = [m]$ for all t then $g_t(\mathbf{w})$ is exactly the objective of Eq. (7), which we denoted by $g(\mathbf{w})$. The convexity of $g(\mathbf{w})$ implies that

$$g\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t\right) \;\le\; \frac{1}{T}\sum_{t=1}^{T} g(\mathbf{w}_t) \;\;.$$

Using the above inequality and Corollary 1, we immediately obtain the following corollary.

Corollary 2 *Assume that the conditions stated in Corollary 1 hold and $I_t = [m]$ for all t and let $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$. Then,*

$$g\left(\bar{\mathbf{w}}\right) \;\le\; g(\mathbf{w}^\star) + \frac{(\sqrt{\sigma}+R)^2\left(1+\log(T)\right)}{2\sigma\,T} \;\;.$$

When $I_t \neq [m]$, Corollary 2 no longer holds. The next theorem bridges this gap as it implies that the same convergence rate still holds in expectation if we randomly choose a stopping time.

Theorem 2 *Assume that the conditions stated in Corollary 1 hold and for all t, I_t is chosen i.i.d. from $[m]$. Additionally, let r be a random index uniformly distributed over $[T]$. Then,*

$$\mathbb{E}_{I_1,\ldots,I_T}\mathbb{E}_r[g(\mathbf{w}_r)] \;\le\; g(\mathbf{w}^\star) + \frac{(\sqrt{\sigma}+R)^2\left(1+\log(T)\right)}{2\sigma\,T} \;\;.$$

For a proof, see [5].

The above theorem implies that, in expectation, the stochastic version of the algorithm would converge as fast as the deterministic version. In the next theorem, whose proof is also given in [5], we provide a concentration bound.

Theorem 3 *Under the assumptions of Thm. 2. Let $\delta \in (0,1)$. Then, with probability of at least $1-\delta$ over the choice of $(I_1,\ldots,I_T)$ and of the index r the following bound holds,*

$$g(\mathbf{w}_r) \;\leq\; g(\mathbf{w}^\star) + \frac{(\sqrt{\sigma}+R)^2\,(1+\log(T))}{2\sigma\,T\,\delta} \quad .$$

Our second concrete derivation is for logistic regression with ℓ_1 regularization. For simplicity, we assume here that the weights of the predcitor we learn are positive. As with the SVM classification problem, we have a set of instance-label pairs $(\mathbf{x}_i, y_i) \in \mathbb{R}^n \times \{-1,+1\}$. The regularized problem on hand now is defined as follows,

$$\min_{\mathbf{w}\in S}\; \sigma\left(\sum_{j=1}^{n} w_j\log(w_j) + \log(n)\right) + \frac{1}{m}\sum_{i=1}^{m}\log\left(1+e^{-y_i\langle\mathbf{w},\mathbf{x}_i\rangle}\right) \quad , \tag{10}$$

where the domain S is defined as follows,

$$S = \left\{\mathbf{w}\,|\,\sum_{j=1}^{n} w_j = 1\,,\;\forall j:\; w_j \geq \epsilon\right\} \quad . \tag{11}$$

The specific algorithm that we obtain for this setting follows the following steps. We start with any vector in S for $\mathbf{w}_1$. In the absence of any prior knowledge, a sensible choice is the vector whose components are all equal to $1/n$. Following the very scheme employed for SVM, on iteration t of the algorithm, we first choose a set $I_t \subseteq [m]$. Then, we replace the objective in Eq. (10) with an instantaneous objective function,

$$g_t(\mathbf{w}) = \sigma\left(\sum_{j=1}^{n} w_j\log(w_j) + \log(n)\right) + \frac{1}{|I_t|}\sum_{i\in I_t}\log\left(1+e^{-y_i\langle\mathbf{w},\mathbf{x}_i\rangle}\right) \quad .$$

We now use Lemma 1 again to get that g_t is σ-strongly convex with respect to $f(\mathbf{w}) = \sum_j w_j\log(w_j) + \log(n)$. As in the case of SVM, we set $\eta_t = 1/(\sigma t)$ while performing the update that matches $f(\mathbf{w})$. From example 4 this update takes the form,

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in S}\sum_{j=1}^{n} w_j\log\left(\frac{w_j}{e^{\theta_j}}\right) \quad , \tag{12}$$

where $\theta_j = \log(w_{t,j}) - \eta_t\lambda_{t,j}$. As in the previous example, $\boldsymbol{\lambda}_t \in \partial g_t(\mathbf{w}_t)$. Note that since $g_t(\mathbf{w})$ is differentiable over S we can simply set,

$$\lambda_{t,j} = \sigma(\log(w_{t,j})+1) - \frac{1}{|I_t|}\sum_{i\in I_t}\frac{y_i x_{i,j}}{1+e^{y_i\langle\mathbf{w}_t,\mathbf{x}_i\rangle}} \quad .$$

From Lemma 8 we know that f is 1-strongly convex with respect to the ℓ_1 norm. Therefore, we need to bound the ℓ_∞ norm of $\boldsymbol{\lambda}_t$ in order to be able to use Thm. 1. The fact that $\mathbf{w}_t \in S$ implies that $|\log(w_{t,j})| \leq \log(1/\epsilon)$. Thus, if we assume that $\|\mathbf{x}_i\|_\infty \leq R$ for all i, we get that,

$$\|\boldsymbol{\lambda}_t\|_\infty \leq \sigma(\log(1/\epsilon)+1) + R \quad .$$

We can now apply Thm. 1 again to obtain the following corollary.

Corollary 3 *Assume that for all $i \in [m]$ the ℓ_∞ norm of $\mathbf{x}_i$ is at most R. Let $\mathbf{w}_1, \ldots, \mathbf{w}_T$ be defined according to Eq. (12) and let $\mathbf{u}$ be an arbitrary vector in S. Then,*

$$\sum_{t=1}^{T} g_t(\mathbf{w}_t) \;\leq\; \sum_{t=1}^{T} g_t(\mathbf{u}) + \frac{(\sigma(\log(1/\epsilon)+1)+R)^2}{2\sigma}\left(1+\log(T)\right) \quad .$$

When $I_t \neq [m]$ we also obtain corollaries analogous to the corollaries obtained for SVM. It remains to show that the update given by Eq. (12) can be computed efficiently. We describe an efficient solution for Eq. (12) in App. B.

A Technical Lemmas and Proofs

The following lemma states that if $\boldsymbol{\lambda} \in \partial f(\mathbf{w})$ then Fenchel-Young inequality holds with equality.

Lemma 5 *Let f be a closed and convex function and let $\partial f(\mathbf{v})$ be its differential set at $\mathbf{v}$. Then, for all $\boldsymbol{\lambda}' \in \partial f(\mathbf{v})$ we have, $f(\mathbf{v}) + f^\star(\boldsymbol{\lambda}') \;=\; \langle \boldsymbol{\lambda}', \mathbf{v} \rangle$.*

Proof Since $\boldsymbol{\lambda}' \in \partial f(\mathbf{v})$, we know that $f(\mathbf{w}) - f(\mathbf{v}) \;\geq\; \langle \boldsymbol{\lambda}', \mathbf{w} - \mathbf{v} \rangle$ for all $\mathbf{w} \in S$. Equivalently

$$\langle \boldsymbol{\lambda}', \mathbf{v} \rangle - f(\mathbf{v}) \;\geq\; \sup_{\mathbf{w} \in S} \left(\langle \boldsymbol{\lambda}', \mathbf{w} \rangle - f(\mathbf{w}) \right) \quad .$$

The right-hand side of the above equals to $f^\star(\boldsymbol{\lambda}')$ and thus,

$$\langle \boldsymbol{\lambda}', \mathbf{v} \rangle - f(\mathbf{v}) \;\geq\; f^\star(\boldsymbol{\lambda}') \quad \Rightarrow \quad \langle \boldsymbol{\lambda}', \mathbf{v} \rangle - f^\star(\boldsymbol{\lambda}') \;\geq\; f(\mathbf{v}) \;. \tag{13}$$

The assumption that f is closed and convex implies that f is the Fenchel conjugate of $f^\star$. Thus,

$$f(\mathbf{v}) = \sup_{\boldsymbol{\lambda}} \left(\langle \boldsymbol{\lambda}, \mathbf{v} \rangle - f^\star(\boldsymbol{\lambda}) \right) \;\geq\; \langle \boldsymbol{\lambda}', \mathbf{v} \rangle - f^\star(\boldsymbol{\lambda}') \;.$$

Combining the above with Eq. (13) gives,

$$\langle \boldsymbol{\lambda}', \mathbf{v} \rangle - f^\star(\boldsymbol{\lambda}') \;\geq\; f(\mathbf{v}) \;\text{ and }\; f(\mathbf{v}) \;\geq\; \langle \boldsymbol{\lambda}', \mathbf{v} \rangle - f^\star(\boldsymbol{\lambda}') \;.$$

Therefore, each of the two inequalities above must hold with equality which concludes the proof. ■

Lemma 6 *Let f be a closed and σ-strongly convex function over S with respect to a norm $\|\cdot\|$. Then, $f^\star$ is differentiable and $\nabla f^\star(\boldsymbol{\theta}) = \arg\max_{\mathbf{w} \in S} \langle \mathbf{w}, \boldsymbol{\theta} \rangle - f(\mathbf{w})$.*

Proof Since f is strongly convex the maximizer of $\max_{\mathbf{w} \in S} \langle \mathbf{w}, \boldsymbol{\theta} \rangle - f(\mathbf{w})$ exists and is unique (see [1] page 19). Denote it by $\pi(\boldsymbol{\theta})$. Since $f^\star$ is a convex function, to prove the lemma it suffices to show that $\partial f^\star(\boldsymbol{\theta}) = \{\pi(\boldsymbol{\theta})\}$. From the definition of $\pi(\boldsymbol{\theta})$ we clearly have that

$$\forall \mathbf{u} \in S, \;\; \langle \pi(\boldsymbol{\theta}), \boldsymbol{\theta} \rangle - f(\pi(\boldsymbol{\theta})) \geq \langle \mathbf{u}, \boldsymbol{\theta} \rangle - f(\mathbf{u}) \;,$$

and thus $\boldsymbol{\theta} \in \partial f(\pi(\boldsymbol{\theta}))$. Therefore, using Lemma 5 we obtain that

$$\langle \pi(\boldsymbol{\theta}), \boldsymbol{\theta} \rangle = f(\pi(\boldsymbol{\theta})) + f^\star(\boldsymbol{\theta}) \;. \tag{14}$$

Let $\boldsymbol{\lambda}$ be an arbitrary vector and to simplify our notation denote $\mathbf{w} = \pi(\boldsymbol{\theta})$ and $\mathbf{u} = \pi(\boldsymbol{\lambda})$. Based on Eq. (14) we have that,

$$\begin{aligned} f^\star(\boldsymbol{\lambda}) - f^\star(\boldsymbol{\theta}) &= \langle \mathbf{u}, \boldsymbol{\lambda} \rangle - f(\mathbf{u}) - \langle \mathbf{w}, \boldsymbol{\theta} \rangle + f(\mathbf{w}) \\ &= f(\mathbf{w}) - f(\mathbf{u}) - \langle \mathbf{w} - \mathbf{u}, \boldsymbol{\lambda} \rangle + \langle \mathbf{w}, \boldsymbol{\lambda} - \boldsymbol{\theta} \rangle \\ &\geq \langle \mathbf{w}, \boldsymbol{\lambda} - \boldsymbol{\theta} \rangle \ , \end{aligned}$$

which implies that $\mathbf{w} \in \partial f^\star(\boldsymbol{\theta})$. Finally, we show that $\mathbf{w}$ is the only element in $\partial f^\star(\boldsymbol{\theta})$ and thus $\mathbf{w} = \nabla f^\star(\boldsymbol{\theta})$. Let $\mathbf{w}_0 \in \partial f^\star(\boldsymbol{\theta})$. Thus $f^\star(\boldsymbol{\theta}) = \langle \mathbf{w}_0, \boldsymbol{\theta} \rangle - f(\mathbf{w}_0) = \langle \mathbf{w}, \boldsymbol{\theta} \rangle - f(\mathbf{w})$. But the uniqueness of the solution of $\max_{\mathbf{w} \in S} \langle \mathbf{w}, \boldsymbol{\theta} \rangle - f(\mathbf{w})$ implies that $\mathbf{w}_0 = \mathbf{w}$. ■

The following lemma yields another criterion for strong convexity.

Lemma 7 *Assume that f is differential. Then f is strongly convex iff*

$$\langle \nabla f(\mathbf{u}) - \nabla f(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \geq \|\mathbf{u} - \mathbf{v}\|^2 \ . \tag{15}$$

Proof Assume f is strongly convex. Then,

$$\begin{aligned} \langle \nabla f(\mathbf{u}), \mathbf{v} - \mathbf{u} \rangle &\leq f(\mathbf{u}) - f(\mathbf{v}) - \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|^2 \\ \langle \nabla f(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle &\leq f(\mathbf{v}) - f(\mathbf{u}) - \frac{1}{2}\|\mathbf{v} - \mathbf{u}\|^2 \end{aligned}$$

Adding these two inequalities we obtain Eq. (15). Assume now that Eq. (15) holds. Define $h(\alpha) = f(\mathbf{v} + \alpha(\mathbf{u} - \mathbf{v}))$, and denote $\mathbf{w} = \mathbf{v} + \alpha(\mathbf{u} - \mathbf{v})$. Then,

$$h'(\alpha) - h'(0) = \langle \nabla f(\mathbf{w}) - \nabla f(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle = \frac{1}{\alpha} \langle \nabla f(\mathbf{w}) - \nabla f(\mathbf{v}), \mathbf{w} - \mathbf{v} \rangle \ .$$

Using Eq. (15) we obtain that

$$h'(\alpha) - h'(0) \geq \frac{1}{\alpha}\|\mathbf{w} - \mathbf{v}\|^2 = \alpha \|\mathbf{u} - \mathbf{v}\|^2 \ .$$

Therefore,

$$\begin{aligned} f(\mathbf{u}) - f(\mathbf{v}) - \langle \nabla f(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle &= h(1) - h(0) - h'(0) \\ &= \int_0^1 (h'(\alpha) - h'(0)) d\alpha \geq \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|^2 \ . \end{aligned}$$

■

Lemma 8 *The function $f(\mathbf{w}) = \sum_{i=1}^n w_i \log(\frac{w_i}{1/n})$ is strongly convex over the probabilistic simplex, $S = \{\mathbf{w} \in \mathbb{R}_+^n : \|\mathbf{w}\|_1 = 1\}$, with respect to the ℓ_1 norm.*

Proof Based on Lemma 7, it suffices to show that for all $\mathbf{u}, \mathbf{v} \in S$ we have

$$\langle \nabla f(\mathbf{u}) - \nabla f(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \geq \|\mathbf{u} - \mathbf{v}\|_1^2 \ . \tag{16}$$

The i'th element of $\nabla f(\mathbf{u})$ is $\log(u_i) + 1 + \log(n)$ and thus we need to show that

$$\sum_i (\log(u_i) - \log(v_i))(u_i - v_i) \geq \|\mathbf{u} - \mathbf{v}\|_1^2 .$$

Let $x_i = (\log(u_i) - \log(v_i))(u_i - v_i)$. Note that the terms $\log(u_i) - \log(v_i)$ and $u_i - v_i$ share the same sign thus $x_i \geq 0$. Denote $I = \{i \in [n] : x_i > 0\}$. Based on the Cauchy-Schwartz inequality, we can bound the right-hand side of the above as follows

$$\begin{aligned}
\|\mathbf{u} - \mathbf{v}\|_1^2 &= \left(\sum_i |u_i - v_i|\right)^2 \leq \left(\sum_{i \in I} \sqrt{x_i} \frac{|u_i - v_i|}{\sqrt{x_i}}\right)^2 \\
&\leq \left(\sum_{j \in I} x_j\right) \left(\sum_{i \in I} \frac{|u_i - v_i|^2}{x_i}\right) \\
&= \left(\sum_{j \in I} (\log(u_j) - \log(v_j))\,(u_j - v_j)\right) \left(\sum_{i \in I} \frac{u_i - v_i}{\log(u_i) - \log(v_i)}\right) \\
&= \langle \nabla f(\mathbf{u}) - \nabla f(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \left(\sum_{i \in I} \frac{u_i - v_i}{\log(u_i) - \log(v_i)}\right)
\end{aligned}$$

.

Therefore, to prove that Eq. (16) holds it suffices to show that

$$\sum_{i \in I} \frac{u_i - v_i}{\log(u_i) - \log(v_i)} \leq 1. \tag{17}$$

To do so, we next show that for all i

$$\frac{u_i - v_i}{\log(u_i) - \log(v_i)} \leq \frac{u_i + v_i}{2} \quad . \tag{18}$$

This inequality immediately implies that Eq. (17) holds since by summing over $i \in I$ we get that the left hand side of Eq. (17) is bounded above by $(\|\mathbf{u}\|_1 + \|\mathbf{v}\|_1)/2 = 1$. The left-hand side of Eq. (18) is positive and thus we can assume without loss of generality that $u_i \geq v_i$. Fix v_i and consider the function

$$\phi(u) = 2(u - v_i) - (u + v_i)(\log(u) - \log(v_i)) .$$

Clearly, $\phi(v_i) = 0$. In addition, the derivative of ϕ is negative,

$$\phi'(u) = 2 - \log(u) - \frac{u + v_i}{u} + \log(v_i) = 1 + \log(\frac{v_i}{u}) - \frac{v_i}{u} \leq 0 ,$$

where to derive the last inequality we used the inequality $\log(a) \leq a - 1$. Therefore, ϕ is monotonically non-increasing and thus $\phi(u) < 0$ in $(v_i, 1]$. We have therefore shown that Eq. (18) holds and our proof is concluded. ■

B Efficient Solution of Eq. (12)

In this section we describe an efficient solution to the following problem,

$$\arg\min_{\mathbf{w}\in S}\sum_{j=1}^{n} w_j \log\left(\frac{w_j}{e^{\theta_j}}\right) \quad \text{where} \quad S=\left\{\mathbf{w}\,|\,\sum_{j=1}^{n} w_j=1\,,\,\forall j:\,w_j\geq\epsilon\right\}\;. \tag{19}$$

For brevity, we denote $u_j = e^{\theta_j}$. We start by writing the Lagrangian of the above constrained optimization problem,

$$\mathcal{L}=\sum_{j=1}^{n} w_j\log(w_j/u_j)+\theta\left(\sum_{j=1}^{n} w_j-1\right)-\sum_{j=1}^{n}\beta_j(w_j-\epsilon)\;,$$

Here, θ is an unconstrained Lagrange multiplier and $\{\beta_j\}$ is a set of *non-negative* Lagrange multipliers for the inequality constraints, $w_j \geq \epsilon$. By taking the derivative of $\mathcal{L}$ with respect to u_i, we get that at the optimum the following should hold,

$$\log(w_i)-\log(u_i)+1+\theta-\beta_i=0\;.$$

After rearranging terms, taking the exponent of the above equation, we can rewrite the above equation as follows,

$$w_i = u_i e^{\beta_i}/Z\;,$$

where $Z = \exp -(\theta+1)$. Since, θ is a Lagrange multiplier for the constraint $\sum_j u_j = 1$ we can also write Z as,

$$Z=\sum_{j=1}^{n} u_j e^{\beta_j}\;.$$

From KKT conditions we know that at the saddle point $(\mathbf{w}^\star,\theta^\star,\{\beta_i^\star\})$ of the Lagrangian $\mathcal{L}$ the following holds,

$$\beta_i^\star(w_i^\star-\epsilon)=0\;.$$

Therefore, if the i'th coordinate of the optimal solution is strictly greater than ϵ we must have that $\beta_i^\star = 0$. In the case where $w_i^\star = \epsilon$ the Lagrange multiplier $\beta_i^\star$ is simply constrained to be non-negative, $\beta_i^\star \geq 0$. Thus, the optimal solution can be rewritten in the following distilled form,

$$w_i^\star=\begin{cases} u_i/Z & w_i^\star>\epsilon \\ \epsilon & \text{otherwise}\end{cases}\;, \tag{20}$$

where Z is set such that $\sum_i w_i^\star = 1$. The lingering question is what components of $\mathbf{u}^\star$ we need to set to ϵ. The following Lemma paves the way to an efficient procedure for determining the components of $\mathbf{w}^\star$ which are equal ϵ.

Lemma 9 *Let $\mathbf{w}^\star$ denote optimal solution of Eq. (19) and let i and j be two indices such that, (i) $u_i \leq u_j$ and (ii) $w_j^\star = \epsilon$, then $w_i^\star = \epsilon$.*

Proof Assume by contradiction that $w_i^\star > \epsilon$. We now use the explicit form of the optimal solution and write,

$$w_i^\star = u_i e^{\beta_i^\star}/Z \;\text{ and }\; w_j^\star = u_j e^{\beta_j^\star}/Z\;.$$

Input:
Sorted vector $\mathbf{u} \in \mathbb{R}_+^n (u_j \leq u_{j+1})$
minimal value constraint $\epsilon > 0$
Initialize:
$Z = \sum_{i=1}^n u_i$
For $l = 1, \ldots, n$:
If $u_l / Z \geq \epsilon$ Then
$l^\star = l - 1$
Break
EndIf
$Z = Z + \frac{\epsilon Z - u_l}{1 - l\epsilon}$
Output: $\mathbf{w}^\star$
$w_j^\star = u_j / Z \quad \text{for} \quad j = l^\star + 1, \ldots, n$
$w_j^\star = \epsilon \quad \text{for} \quad j = 1, \ldots, l^\star$

Figure 1: Pseudo-code of the efficient projection algorithm.

Since $w_j^\star = \epsilon$ we get that $\beta_j^\star \geq 0$ while the fact that $w_i^\star > \epsilon$ implies that $\beta_i^\star = 0$. (See also Eq. (20).) Combining these two facts we get that,

$$w_i^\star = u_i / Z \leq u_j / Z \leq u_j e^{\beta_j^\star} / Z = w_j^\star = \epsilon \ .$$

Thus, $w_i^\star \leq \epsilon$ which stands in contradiction to the assumption $w_i^\star > \epsilon$. ■

The above lemma implies that we if sort the elements of $\mathbf{u}$ in an ascending order, then there exists and index $l^\star$ such that the optimal solution $\mathbf{w}^\star$ takes the following form,

$$\mathbf{w}^\star = (\epsilon, \ldots, \epsilon, \frac{u_{l^\star+1}}{Z}, \frac{u_{l^\star+2}}{Z}, \ldots, \frac{u_n}{Z}) \ .$$

We are therefore left with the task of finding $l^\star$. We do so by examining all possible values for $l \in \{1, \ldots, n\}$. Given a candidate value for $l^\star$, denoted l, we need to check the feasibility of the solution induced by the assumption $l^\star = l$. We next calculate Z for this choice of l. Since the role of Z is to enforce the constraint $\sum_j w_j^\star = 1$ we get that,

$$Z = \frac{\sum_{j=l+1}^n u_j}{1 - l\epsilon} \ .$$

If l indeed induces a feasible solution then for all $j > l$ we must have that $u_j / Z > \epsilon$. Since we assume that the vector $\mathbf{u}$ is sorted in an ascending order it is enough to verify that $u_{l+1} / Z > \epsilon$. In case that we find more than a single feasible solution, it is easy to verify that the smallest candidate for $l^\star$ should be taken. Last, we would like to note that the condition $u_{l+1} / Z > \epsilon$ can be efficiently calculated if we simply keep track of partial sums. Each partial sum is of the form $\sum_{j=l+1}^n u_j$ and is computed from its predecessor $\sum_{j=l}^n u_j$. The pseudo-code describing the entire algorithm is given in Fig.1.

References

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

[2] E. Hazan, A. Kalai, S. Kale, and A. Agarwal. Logarithmic regret algorithms for online convex optimization. In *COLT*, 2006.

[3] J. Kivinen and M. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–64, January 1997.

[4] R.T. Rockafellar and R.J.B Wets. *Variational Analysis*. Springer, New York, 1998.

[5] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *ICML*, 2007.