1 Introduction to convexity

Convexity is an important property that imposes a lot of structure on sets and functions that is particularly useful for online learning and optimization.

1.1 Definitions

We say a set $S \subseteq \mathbb{R}^d$ is convex if for any two points $x, x' \in S$, the line segment $\text{conv}\{x, x'\} := \{(1-\alpha)x + \alpha x' : \alpha \in [0, 1]\}$ between $x$ and $x'$ (also called the convex hull of $\{x, x'\}$) is contained in $S$.

Overloading terms, we say a function $f : S \rightarrow \mathbb{R}$ is convex if its epigraph $\text{epi}(f) := \{(x, t) \in S \times \mathbb{R} : f(x) \leq t\}$ is a convex set (in $\mathbb{R}^d \times \mathbb{R}$).

**Proposition 1.** A function $f : S \rightarrow \mathbb{R}$ is convex if and only if

$$f((1-\alpha)x + \alpha x') \leq (1-\alpha)f(x) + \alpha f(x') \quad \forall x, x' \in S, \alpha \in [0, 1]. \quad (1)$$

Recall that $\text{cl}(S)$ is the closure of $S$ (i.e., the union of $S$ and all of its limit points), $\text{int}(S)$ is the interior of $S$ (i.e., the set of points in $S$ whose small-enough neighborhoods are in $S$), and $\text{bd}(S) := \text{cl}(S) \setminus \text{int}(S)$ is the boundary of $S$.

1.2 Properties of convex sets and functions

A defining property of convex functions is that they have affine approximations to the function at every point that lie entirely below the function. Such affine approximations of a function are given by the subgradients of the function at any point, which are guaranteed to exist. These subgradients can be constructed using the convexity of the function’s epigraph via its supporting hyperplanes.

**Theorem 1.** Let $S \subseteq \mathbb{R}^d$ be convex. For any point $x \in \text{bd}(S)$, there exists $a \in \mathbb{R}^d \setminus \{0\}$ (a normal for a supporting hyperplane) such that

$$\langle a, x' \rangle \leq \langle a, x \rangle \quad \forall x' \in S.$$
We say \( \lambda \in \mathbb{R}^d \) is a subgradient of a function \( f: S \to \mathbb{R} \) at the point \( x \in S \) if
\[
f(x') \geq f(x) + \langle \lambda, x' - x \rangle \quad \forall x' \in S.
\]
The set of subgradients of \( f \) at \( x \) is denoted by \( \partial f(x) \).

**Theorem 2.** Let a convex set \( S \subseteq \mathbb{R}^d \) and a function \( f: S \to \mathbb{R} \) be given.

1. If the subgradient set \( \partial f(x) \) is non-empty for all \( x \in S \), then \( f \) is convex.

2. If \( f \) is convex, then the subgradient set \( \partial f(x) \) is non-empty for all \( x \in \text{int}(S) \).

**Proof.** For the first claim, assume \( \partial f(x) \neq \emptyset \) for all \( x \in S \). Take any \( x, x' \in S \) and \( \alpha \in [0, 1] \). Also take \( \lambda \in \partial f((1-\alpha)x + \alpha x') \), so that
\[
f(x) \geq f((1-\alpha)x + \alpha x') + \langle \lambda, x - ((1-\alpha)x + \alpha x') \rangle
\]
\[
= f((1-\alpha)x + \alpha x') + \alpha \langle \lambda, x - x' \rangle, \tag{1}
\]
\[
f(x') \geq f((1-\alpha)x + \alpha x') + \langle \lambda, x' - ((1-\alpha)x + \alpha x') \rangle
\]
\[
= f((1-\alpha)x + \alpha x') + (1-\alpha) \langle \lambda, x' - x \rangle. \tag{2}
\]
Adding \((1-\alpha)\) times the first inequality and \(\alpha\) times the second inequality gives
\[
(1-\alpha)f(x) + \alpha f(x') \geq f((1-\alpha)x + \alpha x').
\]
Since the above inequality holds for all choices of \( x, x' \in S \) and \( \alpha \in [0, 1] \), Proposition \ref{prop:convexity} implies that \( f \) is convex.

For the second claim, assume \( f \) is convex, pick any \( x \in \text{int}(S) \), and consider the point \( (x, f(x)) \in \text{bd}(\text{epi}(f)) \). By Theorem \ref{thm:subgradients} there exists \((a, b) \in (\mathbb{R}^d \times \mathbb{R}) \setminus \{(0, 0)\}\) such that
\[
\langle a, x' \rangle + bt \leq \langle a, x \rangle + bf(x) \quad \forall (x', t) \in \text{epi}(f).
\]
In particular, this holds for \((x', t) = (x, f(x) + \epsilon) \in \text{epi}(f)\) for any \( \epsilon > 0 \), so \( be \leq 0 \); this implies \( b \leq 0 \). If \( b = 0 \), then \( \langle a, x' - x \rangle \leq 0 \) for all \( x' \in S \), which is impossible because \( a \neq 0 \) and \( x \in \text{int}(S) \). Thus, it must be that \( b < 0 \), so
\[
\langle a/b, x' \rangle + f(x') \geq \langle a/b, x \rangle + f(x) \quad \forall x' \in S,
\]
which rearranges to
\[
f(x') \geq f(x) + \langle -a/b, x' - x \rangle \quad \forall x' \in S.
\]
This implies that \( \lambda := -a/b \in \partial f(x) \), so \( \partial f(x) \neq \emptyset \). \( \square \)
Figure 1: The graph of the function $f(x) = x^2$ is depicted in red. The (sub)gradient of $f$ at $x = 1/2$ is $\lambda := 1$. The supporting hyperplane of $\text{epi}(f)$ at $(1/2, 1/4)$, as well as its normal vector $(\lambda, -1)$ (shown translated to $(1/2, 1/4)$), are depicted in blue.

For a geometric picture of a subgradient $\lambda \in \partial f(x)$, observe that $(\lambda, -1)$ defines a supporting hyperplane for $\text{epi}(f)$ at $(x, f(x))$.

**Proposition 2.** If $f: S \to \mathbb{R}$ is convex and differentiable at $x$, then the gradient of $f$ at $x$ is a subgradient of $f$ at $x$: $\nabla f(x) \in \partial f(x)$.

**Proof.** By Proposition 1 the convexity of $f$ implies that for any $x, x' \in S$ and $\alpha \in (0, 1)$,

$$
\alpha f(x') \geq f((1 - \alpha)x + \alpha x') - f(x) = \alpha f(x) + f((1 - \alpha)x + \alpha x') - f(x).
$$

Dividing both sides by $\alpha$ and taking the limit as $\alpha \to 0$,

$$
f(x') \geq f(x) + \lim_{\alpha \to 0} \frac{f(x + \alpha(x' - x)) - f(x)}{\alpha} = f(x) + \langle \nabla f(x), x' - x \rangle.
$$

Thus $\nabla f(x) \in \partial f(x)$. \qed

Henceforth, we shall only consider convex functions whose epigraphs are closed sets and whose domains are non-empty (unless otherwise specified). Such convex functions are said to be closed and proper. This rules out pathological functions that we choose to ignore in our applications.

There are a few simple but useful operations that preserve convexity. We give these operations below and describe their subgradient sets. Below, $T$ is an abstract set; and $g, g_1, g_2, \ldots, g_n$, and $g_t$ for $t \in T$ are assumed to be convex.
Addition. If \( f(x) = g_1(x) + g_2(x) \), then \( \partial f = \partial g_1 + \partial g_2 \).

Positive scaling. If \( f(x) = \alpha \cdot g(x) \) for some \( \alpha > 0 \), then \( \partial f = \alpha \cdot \partial g \).

Affine composition. If \( f(x) = g(Ax + b) \), then \( \partial f(x) = A^\top \partial g(Ax + b) \).

Finite pointwise maximum. If \( f(x) = \max_{i \in [n]} g_i(x) \), then
\[
\partial f(x) = \text{conv} \left( \bigcup_{i \in [n]: g_i(x) = f(x)} \partial g_i(x) \right).
\]

Arbitrary pointwise supremum. If \( f(x) = \sup_{t \in \mathcal{T}} g_t(x) \), then
\[
\partial f(x) = \text{conv} \left( \bigcup_{t \in \mathcal{T}: g_t(x) = f(x)} \partial g_t(x) \right).
\]

Caveat: the definition of \( \partial f \) only holds if some “regularity conditions” on \( g_t \) and \( \mathcal{T} \) hold (e.g., \( \mathcal{T} \) is compact and \( t \mapsto g_t(x) \) is continuous for each \( x \)).

More general composition. If \( h : \mathbb{R}^n \to \mathbb{R} \) is convex and non-decreasing in each argument, and \( f(x) := h(g_1(x), g_2(x), \ldots, g_n(x)) \), then
\[
\partial f(x) = \bigcup_{(\lambda_1, \lambda_2, \ldots, \lambda_n) \in \partial h(g_1(x), g_2(x), \ldots, g_n(x))} \partial \left( \sum_{i=1}^n \lambda_i g_i(x) \right).
\]

1.3 Convex optimization

Convex functions are special because information about the function at any given point provides information about the entire function. For example, the existence of subgradients guaranteed by Theorem 2 implies that an affine approximation at any given point provides lower bounds on the function everywhere.

Another piece of local information that has global implications is local optimality. We say a point \( x \in \text{int}(S) \) is a local minimizer of a function \( f : S \to \mathbb{R} \) if there exists \( \delta > 0 \) such that \( f(x) \leq f(x + v) \) for all \( v \in \mathbb{R}^d \) with \( \|v\| \leq \delta \). We say a point \( x \in S \) is a (global) minimizer of \( f \) if \( f(x) \leq f(x') \) for all \( x' \in S \). We first describe this local-to-global phenomenon when the domain of \( f \) is all of \( \mathbb{R}^d \).
Proposition 3. Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex. If $x \in \mathbb{R}^d$ is a local minimizer of $f$, then $x$ is a (global) minimizer of $f$. Moreover, $x$ is a minimizer of $f$ if and only if $0 \in \partial f(x)$.

Proof. Suppose $x$ is a local minimizer of $f$. Take $\delta > 0$ such that $f(x) \leq f(x + v)$ for all $v \in \mathbb{R}^d$ with $\|v\| \leq \delta$. Now consider any $x' \in S$. We'll show that $f(x) \leq f(x')$. We may assume that $\|x - x'\| \geq \delta$. Let $\alpha := \delta/\|x' - x\|$ so that $\|\alpha(x' - x)\| \leq \delta$. Then $f(x) \leq f(x + \alpha(x' - x)) \leq (1 - \alpha)f(x) + \alpha f(x')$ by Proposition 1. Rearranging, this reads $f(x) \leq f(x')$.

By definition, $x$ is a minimizer if and only if

$$f(x') \geq f(x) = f(x) + \langle 0, x' - x \rangle \quad \forall x' \in \mathbb{R}^d,$$

which is precisely equivalent to $0 \in \partial f(x)$. \qed

The second part of Proposition 3 gives an optimality condition for unconstrained convex optimization: a point $x$ is a minimizer of $f$ if and only if $0$ is among the subgradients of $f$ at $x$.

When the domain of $f$ is a closed convex $S \subseteq \mathbb{R}^d$, the optimality condition is somewhat more involved.

Proposition 4. Let $S \subseteq \mathbb{R}^d$ be a closed convex set, and let a convex function $f : S \to \mathbb{R}$ be given. The point $x \in S$ is a minimizer of $f$ if and only if there exists $\lambda \in \partial f(x)$ such that

$$\langle \lambda, x - x' \rangle \leq 0 \quad \forall x' \in S.$$

To prove Proposition 4, we need to discuss extended convex functions, which are convex functions $f : \mathbb{R}^d \to \mathbb{R}$, where $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$. We can extend a convex function $f : S \to \mathbb{R}$ so that it is defined over all of $\mathbb{R}^d$ by letting $f(x) := +\infty$ for $x \notin S$. In other words, we consider the function $f + I_S$, where $I_S : \mathbb{R}^d \to \overline{\mathbb{R}}$ is the set indicator function for $S$, given by

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

All of the properties of ($\mathbb{R}$-valued) convex functions discussed above also apply to extended convex functions. Moreover, the set indicator function $I_S$ for any (closed) convex set $S$ is a (closed) extended convex function, so $f + I_S$ is also a (closed) extended convex function.

Proposition 4 turns out to simply be a restatement of the optimality condition from Proposition 3, namely that $0 \in \partial (f + I_S)(x)$. This is equivalent
to the existence of a subgradient $\lambda \in \partial f(x)$ such that $-\lambda \in \partial I_S(x)$. All that remains to do is to characterize $\partial I_S(x)$. It turns out $\partial I_S(x)$ is simply the set of normal vectors for (possibly degenerate) supporting hyperplanes of $S$ at $x$. This set is called the normal cone $N_S(x)$ for $S$ at $x$:

$$N_S(x) := \{a \in \mathbb{R}^d : \langle a, x' \rangle \leq \langle a, x \rangle \ \forall x' \in S\}.$$  

So $\partial I_S(x) = N_S(x)$.

2 Online convex optimization

The most basic online convex optimization setting we shall deal with is the following.

**Algorithm 1** Protocol for online convex optimization

<table>
<thead>
<tr>
<th>input</th>
<th>Closed convex set $S \subseteq \mathbb{R}^d$.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: for $t = 1$ to $T$ do</td>
<td></td>
</tr>
<tr>
<td>2: Choose decision vector: $w_t \in S$.</td>
<td></td>
</tr>
<tr>
<td>3: Receive feedback about convex function $f_t: \mathbb{R}^d \rightarrow \mathbb{R}$.</td>
<td></td>
</tr>
<tr>
<td>4: Incur loss: $f_t(w_t)$.</td>
<td></td>
</tr>
<tr>
<td>5: end for</td>
<td></td>
</tr>
</tbody>
</table>

The total loss incurred by the learner is $L_T := \sum_{t=1}^{T} f_t(w_t)$. The total loss incurred by any fixed vector $u \in S$ is $L_{T,u} := \sum_{t=1}^{T} f_t(u)$. The regret of the learner to the decision point $u$ is $R_{T,u} := L_T - L_{T,u}$, and the regret of the learner (to the best decision point $u \in S$ in hindsight) is $R_T := L_T - \min_{u \in S} L_{T,u}$. Sometimes, we will consider the regret of the learner to the best decision point $u \in U$ in hindsight, for other set $U \subseteq S$ possibly different from $S$ (especially when $S = \mathbb{R}^d$). We denote this quantity by $R_{T,U} := L_T - \min_{u \in U} L_{T,u}$.

We have seen special cases of this problem before (e.g., the linear loss game, where $f_t$ are linear functions; linear binary prediction with squared hinge loss). In the case of online binary prediction, there is some side-information that comes before the learner chooses $w_t$, namely the feature vector $x_t$. But in that problem, the correct label $y_t$ is not revealed until after $w_t$ is chosen, and both $y_t$ and $x_t$ are essential parts of the loss function $f_t$ (e.g., for squared hinge loss $h_l(w) = \max\{0,1 - y_t\langle x_t, w \rangle\}^2$). More generally, we will stipulate what information is provided to the learner about the function $f_t$. 


2.1 Follow the leader

Since the loss functions $f_t$ are all convex, so is their sum. So, in particular, $L_{T,u}$ is a convex function of $u \in S$. Because of the special structure afforded by convexity, it is typically possible to minimize $L_{T,u}$ over $u \in S$ (assuming reasonable access to a description of $S$) in a computationally efficient manner. This is similar to the surrogate loss bounds used in the analyses of Perceptron and Winnow, where the surrogate loss function was the hinge loss $h_l$, a convex function. (A difference is that the surrogate loss bounds we considered were sums only over rounds in which the learner makes a mistake; here, $L_{T,u}$ is a sum over all rounds from 1 to $T$.) Ultimately, this means that the offline convex optimization problem is computationally tractable.

The first algorithm to consider is called “Follow the leader”. In round $t$, the learner simply solves the offline convex optimization problem of minimizing $\sum_{i=1}^{t-1} f_i$ over $u \in S$. This means that the learner must have received information about the functions $f_1, f_2, \ldots, f_{t-1}$ sufficient to find a minimizer of their sum.

**Algorithm 2** Follow the leader

\begin{align*}
\textbf{input} & \quad \text{Closed convex set } S \subseteq \mathbb{R}^d. \\
1: & \quad \text{Pick } w_1 \in S. \\
2: & \quad \text{for } t = 1 \text{ to } T \text{ do} \\
3: & \quad \text{Receive feedback about convex function } f_t : \mathbb{R}^d \to \mathbb{R}. \\
4: & \quad \text{Incur loss: } f_t(w_t). \\
5: & \quad \text{Choose next vector: } w_{t+1} \in \arg\min_{u \in S} \sum_{i=1}^{t} f_i(u). \\
6: & \quad \text{end for}
\end{align*}

2.2 Quadratic losses

We’ll first consider Algorithm 2 in the context where every loss function $f_t$ is of the form

$$f_t(w) = \frac{1}{2} \|w - x_t\|^2$$

for some $x_t \in S$. This can be interpreted as the negative log-likelihood of the point $x_t$ under the isotropic Gaussian distribution $N(w, I)$ (up to some additive constant that does not depend on $w$).

In this case, we can analytically determine the minimizer of $\sum_{i=1}^{t} f_i$ over all of $\mathbb{R}^d$ simply by finding $w \in \mathbb{R}^d$ such that $0 \in \partial (\sum_{i=1}^{t} f_i(w))$ (as per Proposition 3). Each function $f_i$ is differentiable at every $w \in \mathbb{R}^d$—the
gradient at $w$ is $\nabla f_i(w) = w - x_i$. Thus, we simply have to solve

$$\sum_{i=1}^{t} \nabla f_i(w) = \sum_{i=1}^{t} (w - x_i) = 0$$

for $w \in \mathbb{R}^d$. The solution is

$$w_{t+1} := \frac{1}{t} \sum_{i=1}^{t} x_i.$$  

Since $w_{t+1}$ is a convex combination of points $x_i$, each of which is in the convex set $S$, it follows that $w_{t+1} \in S$. Therefore, $w_{t+1}$ happens to also minimize $\sum_{i=1}^{t} f_i$ over $S$.

Because all of the $f_i$ are quadratic, it is possible to characterize the regret of Algorithm 2 in nearly closed-form. We shall not do this, but instead take a different route.

**Lemma 1** (Be the leader). Let $g_1, g_2, \ldots$ be any sequence of real-valued functions with a common domain $S \subseteq \mathbb{R}^d$. Assume $w_t \in \arg\min_{u \in S} \sum_{i=1}^{t-1} g_i(u)$ for all $t \in \mathbb{N}$. For all $T \in \mathbb{N}$,

$$\sum_{t=1}^{T} g_t(w_t) - g_t(u) \leq \sum_{t=1}^{T} g_t(w_t) - g_t(w_{t+1}) \quad \text{for all } u \in S.$$  

**Remark 1.** We call this the “Be the leader” lemma because it shows that a learner who, gets access to $g_t$ before having to make a decision in round $t$, has zero regret. That is, if in round $t$, the learner uses $\arg\min_{w \in S} \sum_{i=1}^{t} g_i(w)$ (which we are calling $w_{t+1}$), then $R_{T,u} \leq 0$ for all $u \in S$.

**Remark 2.** Note that this lemma does not require the functions $g_t$ to be convex, nor does it require the domain $S$ to be convex.

**Proof.** We simply have to prove

$$\sum_{t=1}^{T} g_t(w_{t+1}) \leq \sum_{t=1}^{T} g_t(u) \quad \text{for all } u \in S.$$  

The proof is by induction on $T$. For $T = 1$, the claim is simply that $g_1(w_2) \leq g_1(u)$ for all $u \in S$, which is true by definition of $w_2$. Now consider any $T > 1$. We have

$$\sum_{t=1}^{T} g_t(w_{t+1}) = g_1(w_2) + \sum_{t=2}^{T} g_t(w_{t+1}) \leq g_1(u) + \sum_{t=2}^{T} g_t(u) = \sum_{t=1}^{T} g_t(u),$$

where the first inequality follows from the inductive hypothesis, and the second inequality follows because $w_{t+1}$ is the optimal decision in round $t$. This completes the proof.
1, and assume (as the inductive hypothesis) \( \sum_{t=1}^{T-1} g_t(w_{t+1}) \leq \sum_{t=1}^{T-1} g_t(u) \) for all \( u \in S \). Then
\[
\sum_{t=1}^{T} g_t(w_{t+1}) = \sum_{t=1}^{T-1} g_t(w_{t+1}) + g_T(w_{T+1}) \\
\leq \sum_{t=1}^{T-1} g_t(w_{T+1}) + g_T(w_{T+1}) \\
\text{by inductive hypothesis, with } u = w_{T+1} \\
= \sum_{t=1}^{T} g_t(w_{T+1}) \\
\leq \sum_{t=1}^{T} g_t(u) \text{ for all } u \in S, \text{ by definition of } w_{T+1}. \]

Therefore, to bound \( R_{T,u} \) for any \( u \in S \), we simply have to bound \( \sum_{t=1}^{T} f_t(w_t) - f_t(w_{t+1}) \), where the \( w_t \) are given in Algorithm 2. This is straightforward in the case where \( f_t(w) = 0.5\|w - x_t\|^2 \), because
\[
w_{t+1} = w_t + \frac{1}{t}(x_t - w_t).
\]

Indeed,
\[
\frac{1}{2}\|w_t - x_t\|^2 - \frac{1}{2}\|w_{t+1} - x_t\|^2 = \frac{1}{2}\|w_t - x_t\|^2 - \frac{1}{2}\left\|w_t - x_t + \frac{1}{t}(x_t - w_t)\right\|^2 \\
= -\frac{1}{t}\langle x_t - w_t, x_t - x_t \rangle - \frac{1}{2t^2}\|x_t - w_t\|^2 \\
= \frac{1}{t}\left(1 - \frac{1}{2t}\right)\|x_t - w_t\|^2 \leq \frac{1}{t}\|x_t - w_t\|^2.
\]

By Lemma 1 and the calculation above, we have the following theorem.

**Theorem 3.** Suppose \( S \) is a closed convex set containing \( 0 \) with diameter at most \( D \) (i.e., \( \sup_{u,v \in S} \|u - v\| \leq D \)). Also, assume every \( f_t: \mathbb{R}^d \rightarrow \mathbb{R} \) is of the form \( f_t(w) = 0.5\|w - x_t\|^2 \) for some \( x_t \in S \). The regret of the learner using Algorithm 2 (“Follow the leader”) with \( w_1 := 0 \) is bounded as
\[
R_{T,u} \leq D^2 \sum_{t=1}^{T} \frac{1}{t} \leq D^2 \log(T + 1).
\]
2.3 Linear losses

Unfortunately, Algorithm 2 does not always enjoy the same regret bound as in Theorem 3 for general online convex optimization. In fact, its failings can be seen even with online linear optimization. Suppose \( d := 1, S := [-1, 1] \), and

\[
f_t(w) := \begin{cases} 
-\frac{1}{2}w & \text{if } t = 1, \\
+w & \text{if } t \text{ is even}, \\
-w & \text{if } t \text{ is odd and } t > 1.
\end{cases}
\]

For any \( t \geq 1 \),

\[
\sum_{i=1}^{t} f_i(w) = \begin{cases} 
-\frac{1}{2}w & \text{if } t \text{ is odd}, \\
+\frac{1}{2}w & \text{if } t \text{ is even}.
\end{cases}
\]

Therefore, Algorithm 2 picks

\[
w_{t+1} = \begin{cases} 
+1 & \text{if } t \text{ is odd}, \\
-1 & \text{if } t \text{ is even}.
\end{cases}
\]

Observe that \( f_t(w_t) = 1 \) for all \( t > 1 \), while \( f_t(0) = 0 \) for all \( t \). Therefore, the regret of a learner using Algorithm 2 is at least \( T - \Theta(1) \).

When \( S \) is a bounded convex set (as in the case of \( S = [-1, 1] \)), then the minimum of a linear function is always attained at a point on the boundary of \( S \), which can change wildly from round to round. This stands in stark contrast to the case where \( f_t(w) := ||w - x_t||^2 \), where the minimizer of \( \sum_{i=1}^{t-1} f_i \) was an average of the points \( x_1, x_2, \ldots, x_{t-1} \) (and hence inside their convex hull).

Note that we have already seen algorithms for special cases of online linear optimization. For example, when \( S = \Delta^{d-1} \) (the probability simplex) and the linear functions are of the form \( w \mapsto \langle \ell_t, w \rangle \) for some \( \ell_t \in \mathbb{R}_+^d \), we saw that the Hedge algorithm provides a sublinear regret bound. The Hedge algorithm chooses its next decision point \( w_{t+1} \) to minimize a sum of all previous linear loss functions \( \sum_{i=1}^{t} \ell_i \), plus another term that encourages \( w_{t+1} \) to be close to \( w_t \).

3 Linearization and regularization

As we have seen in the examples above, linear functions seem to cause problems for algorithms such as Algorithm 2. In some sense, linear (or affine) functions are the worst convex functions for any online convex optimization algorithm.
To make the previous assertion a bit more precise, consider a specialization of the protocol in Algorithm 1 where the feedback provided to the learner is simply \( f_t(w_t) \) (the value of the convex function), and an arbitrary subgradient \( \lambda_t \in \partial f_t(w_t) \). From the point of view of the learner, the function \( f_t \) might very well have been the affine function \( w \mapsto f_t(w_t) + \langle \lambda_t, w - w_t \rangle \).

Moreover, for any \( u \in S \),
\[
R_{T,u} = \sum_{t=1}^{T} (f_t(w_t) - f_t(u)) \leq \sum_{t=1}^{T} \langle \lambda_t, w_t - u \rangle,
\]
which follows from the fact \( \lambda_t \in \partial f_t(w_t) \) and definition of subgradients. For now, we shall continue with this feedback model.

Taking inspiration from Hedge, we shall modify Algorithm 2 in a way that encourages \( w_{t+1} \) to be close to \( w_t \). One way to do this is to minimize a sum of the convex functions \( f_1, f_2, \ldots, f_t \) (or affine approximations thereof), plus a convex regularization term that promotes stability. For now, we simply use the term \( \|w\|^2/(2\eta) \) for some \( \eta > 0 \). This is a penalty for \( w \) with large norm, with the strength of the penalty being controlled by \( \eta \). Therefore, all \( w_t \) are encouraged to be close to the zero vector, and hence can not change too much from round to round. We call the resulting algorithm “Follow the regularized leader” (using linearized losses), shown in Algorithm 3.

---

**Algorithm 3** Follow the regularized leader

**input** Closed convex set \( S \subseteq \mathbb{R}^d \), regularization parameter \( \eta > 0 \).

1. Let \( w_1 := \arg \min_{u \in S} \frac{1}{2\eta} \|u\|^2 \).
2. for \( t = 1 \) to \( T \) do
3.  receive feedback: \( f_t(w_t) \) and \( \lambda_t \in \partial f_t(w_t) \) for some convex function \( f_t: \mathbb{R}^d \to \mathbb{R} \).
4.  incur loss: \( f_t(w_t) \).
5.  update: \( w_{t+1} := \arg \min_{u \in S} \sum_{i=1}^{t} \langle \lambda_i, u \rangle + \frac{1}{2\eta} \|u\|^2 \). (2)
6. end for

---

### 3.1 Unconstrained domain

When \( S = \mathbb{R}^d \), we have a closed-form expression for \( w_{t+1} \) as given in (2), since the gradient of \( \sum_{i=1}^{t} (\lambda_i, u) + \|u\|^2/(2\eta) \) is \( \sum_{i=1}^{t} \lambda_i + u/\eta \), which is
equal to zero at the point

\[ w_{t+1} := -\eta \sum_{i=1}^{t} \lambda_i. \]

Observe that we have the recursive definition of \( w_{t+1} \):

\[ w_{t+1} = w_t - \eta \lambda_t, \]

so \( \eta > 0 \) can be interpreted as a step size (or learning rate). We call this algorithm “Online gradient descent” (though maybe it should be called “Online subgradient descent”).

**Algorithm 4 Online gradient descent**

**input** Step size \( \eta > 0 \).

1: Let \( w_1 := 0 \).
2: for \( t = 1 \) to \( T \) do
3: \hspace{1em} Receive feedback: \( f_t(w_t) \) and \( \lambda_t \in \partial f_t(w_t) \) for some convex function \( f_t: \mathbb{R}^d \to \mathbb{R} \).
4: \hspace{1em} Incur loss: \( f_t(w_t) \).
5: \hspace{1em} Update: \( w_{t+1} := w_t - \eta \lambda_t \).
6: end for

**Theorem 4.** Let \( f_1, f_2, \ldots, f_T: \mathbb{R}^d \to \mathbb{R} \) be any sequence of convex loss functions. The total loss incurred by a learner using Algorithm 4 (“Online gradient descent”) with step size \( \eta > 0 \) is bounded as

\[
\sum_{t=1}^{T} f_t(w_t) \leq \sum_{t=1}^{T} f_t(u) + \frac{1}{2\eta} ||u||^2 + \eta \sum_{t=1}^{T} ||\lambda_t||^2
\]

for any \( u \in \mathbb{R}^d \).

**Proof.** Fix any \( u \in \mathbb{R}^d \). Since \( \lambda_t \in \partial f_t(w_t) \) for all \( t = 1, 2, \ldots, T \),

\[
f_t(u) - f_t(w_t) \geq \langle \lambda_t, u - w_t \rangle
\]

by the definition of subgradients. This implies

\[
\sum_{t=1}^{T} (f_t(w_t) - f_t(u)) \leq \sum_{t=1}^{T} \langle \lambda_t, w_t - u \rangle.
\] (3)
Define $g_0(w) := \|w\|^2/(2\eta)$, and $g_t(w) := \langle \lambda_t, w \rangle$ for $t = 1, 2, \ldots, T$. By Lemma 1 (with index $t$ starting at zero), using the fact that $w_t := \arg\min_{w \in \mathbb{R}^d} \sum_{i=0}^{t-1} g_i(w)$ $= \arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^{t-1} \langle \lambda_t, w \rangle + \|w\|^2/(2\eta)$ (even for $t = 1$),

$$
\sum_{t=0}^{T} g_t(w_t) - g_t(u) \leq \sum_{t=0}^{T} g_t(w_t) - g_t(w_{t+1})
$$

for all $u \in \mathbb{R}^d$ (where $w_0 \in \mathbb{R}^d$ is arbitrary). The left-hand side is equal to

$$
g_0(w_1) - g_0(u) + \sum_{t=1}^{T} \langle \lambda_t, w_t - u \rangle,
$$

and the right-hand side is equal to

$$
g_0(w_0) - g_0(w_1) + \sum_{t=1}^{T} \langle \lambda_t, w_t - w_{t+1} \rangle.
$$

Rearranging gives

$$
\sum_{t=1}^{T} \langle \lambda_t, w_t - u \rangle \leq g_0(u) - g_0(w_1) + \sum_{t=1}^{T} \langle \lambda_t, w_t - w_{t+1} \rangle.
$$

Combining this with (3) gives

$$
\sum_{t=1}^{T} (f_t(w_t) - f_t(u)) \leq g_0(u) - g_0(w_1) + \sum_{t=1}^{T} \langle \lambda_t, w_t - w_{t+1} \rangle.
$$

Now, using the fact that $w_{t+1} = w_t - \eta \lambda_t$, we have

$$
\sum_{t=1}^{T} \langle \lambda_t, w_t - w_{t+1} \rangle = \sum_{t=1}^{T} \langle \lambda_t, \eta \lambda_t \rangle \leq \eta \sum_{t=1}^{T} \| \lambda_t \|^2.
$$

Plugging back into (4) completes the proof.

\[ \square \]

**Corollary 1.** Let $f_1, f_2, \ldots, f_T: \mathbb{R}^d \to \mathbb{R}$ be any sequence of convex loss functions. Let $U := \{ u \in \mathbb{R}^d : \|u\| \leq B \}$, and assume $G > 0$ is such that $\sum_{t=1}^{T} \| \lambda_t \|^2 \leq G^2T$, where $\lambda_t \in \partial f_t(w_t)$ for each $t = 1, 2, \ldots, T$. The regret $\mathcal{R}_{T,U}$ to the best decision point $u \in U$ incurred by a learner using Algorithm 4 using step size $\eta := B/(G\sqrt{2T})$ is bounded as

$$
\mathcal{R}_{T,U} \leq BG\sqrt{2T}.
$$
3.2 Constrained domain

In general, we may be interested in a general convex domain \( S \subseteq \mathbb{R}^d \). In this case, it is possible that online gradient descent will step outside of \( S \): we may have \( w_t - \eta \lambda_t \notin S \) even if \( w_t \in S \). Fortunately, the update rule (2) can be expressed as a Euclidean projection of the scaled sum of negative subgradients onto the set \( S \):

\[
\begin{align*}
m_{t+1} &:= -\eta \sum_{i=1}^{t} \lambda_i, \\
w_{t+1} &:= \Pi_S (m_{t+1}) := \arg\min_{w \in S} \| w - m_{t+1} \|^2. 
\end{align*}
\]

Therefore, in Algorithm 3 it suffices to accumulate the negative sum of subgradients, and to compute Euclidean projections to \( S \). (A suitable description of \( S \) must be provided so that such projections can be computed.)

**Theorem 5.** Let \( f_1, f_2, \ldots, f_T : \mathbb{R}^d \to \mathbb{R} \) be any sequence of convex loss functions, and \( S \subseteq \mathbb{R}^d \) a closed convex set. The total loss incurred by a learner using Algorithm 3 (“Follow the regularized leader”) with regularization parameter \( \eta > 0 \) is bounded as

\[
\sum_{t=1}^{T} f_t(w_t) \leq \sum_{t=1}^{T} f_t(u) + \frac{1}{2\eta} \|u\|^2 + 2\eta \sum_{t=1}^{T} \| \lambda_t \|^2 \quad \text{for all } u \in S.
\]

**Proof.** We follow the proof of Theorem 4 up to the step in (4) (for an arbitrary \( u \in S \)):

\[
\sum_{t=1}^{T} (f_t(w_t) - f_t(u)) \leq g_0(u) - g_0(w_1) + \sum_{t=1}^{T} \langle \lambda_t, w_t - w_{t+1} \rangle. \quad (6)
\]

It remains to bound \( \langle \lambda_t, w_t - w_{t+1} \rangle \) for \( t = 1, 2, \ldots, T \). Define

\[
h_t(w) := \eta \sum_{i=1}^{t} \lambda_i, w + \frac{1}{2} \| w \|^2
\]

so

\[
w_t = \arg\min_{w \in S} h_{t-1}(w) \quad \text{for all } t = 1, 2, \ldots, T + 1.
\]

Since \( h_t \) is a quadratic function, it is straightforward to check that

\[
h_t(w_t) = h_t(w_{t+1}) + \langle \nabla h_t(w_{t+1}), w_t - w_{t+1} \rangle + \frac{1}{2} \| w_t - w_{t+1} \|^2
\]
(the second-order Taylor expansion of $h_t$ around $w_{t+1}$ at $w_t$). Therefore,

$$h_t(w_{t+1}) - h_t(w_t) = \langle \nabla h_t(w_{t+1}), w_{t+1} - w_t \rangle - \frac{1}{2} \| w_t - w_{t+1} \|^2$$

$$\leq -\frac{1}{2} \| w_t - w_{t+1} \|^2,$$

where the inequality follows from the optimality condition in Proposition 4 and the fact that $w_{t+1}$ minimizes $h_t(w)$ over all $w \in S$. Moreover,

$$h_t(w_{t+1}) - h_t(w_t) = h_{t-1}(w_{t+1}) - h_{t-1}(w_t) + \eta(\lambda_t, w_{t+1} - w_t)$$

$$\geq \eta(\lambda_t, w_{t+1} - w_t)$$

since $w_t$ minimizes $h_{t-1}(w)$ over all $w \in S$. Combining the previous two displayed inequalities gives

$$\frac{1}{2} \| w_t - w_{t+1} \|^2 \leq \eta(\lambda_t, w_t - w_{t+1}).$$

By Cauchy-Schwarz,

$$\eta(\lambda_t, w_t - w_{t+1}) \leq \eta \| \lambda_t \| \| w_t - w_{t+1} \|.$$

This means that $\| w_t - w_{t+1} \| \leq 2 \eta \| \lambda_t \|$, and also

$$\langle \lambda_t, w_t - w_{t+1} \rangle \leq \| \lambda_t \| \| w_t - w_{t+1} \| \leq 2 \eta \| \lambda_t \|^2.$$

Plugging back into (6) completes the proof.

Corollary 2. Let $f_1, f_2, \ldots, f_T : \mathbb{R}^d \to \mathbb{R}$ be any sequence of convex loss functions, and $S \subseteq \mathbb{R}^d$ a closed convex set. Let $U := \{ u \in S : \| u \| \leq B \}$, and assume $G > 0$ is such that $\sum_{t=1}^T \| \lambda_t \|^2 \leq G^2 T$, where $\lambda_t \in \partial f_t(w_t)$ for each $t = 1, 2, \ldots, T$. The regret $R_{T,U}$ to the best decision point $u \in U$ incurred by a learner using Algorithm 3 using regularization parameter $\eta := B/(G\sqrt{T})$ is bounded as

$$R_{T,U} \leq 2BG\sqrt{T}.$$

If $B$ is such that $\max_{u \in S} \| u \| \leq B$, then $R_T \leq 2BG\sqrt{T}$.

Example 1. Suppose $S = \mathbb{R}_+^d$ (the non-negative orthant) and $f_t(w) := \max\{0, 1 - y_t(\langle x_t, w \rangle - \theta)\}$ for $x_t \in \mathbb{R}_+^d$, $y_t \in \{\pm 1\}$, and $\theta \geq 0$. For this example, suppose the feedback provided is $(x_t, y_t) \in \mathbb{R}_+^d \times \{\pm 1\}$. Then, to implement Algorithm 3 we may let $\lambda_t \in \partial f_t(w_t)$ be any subgradient. The most natural choice is

$$\lambda_t := \begin{cases} 0 & \text{if } y_t(\langle x_t, w_t \rangle - \theta) \geq 1 \\ -y_t x_t & \text{otherwise.} \end{cases}$$
So accumulate $m_{t+1} := -\eta \sum_{i=1}^t \lambda_i = \eta \sum_{i\in[t]:y_i((x_i,w_t)-\theta) < 1} y_i x_i$, and set $w_{t+1} := \arg\min_{w\in\mathbb{R}^d} \|w - m_{t+1}\|^2$, which is simply obtained by setting any negative entries in $m_{t+1}$ to zero.

It is also possible to modify online gradient descent so that it always picks decision points in $S$, simply by using the update rule

$$w_{t+1} := \arg\min_{w\in S} \|w - (w_t - \eta \lambda_t)\|^2.$$ 

This modification (Algorithm 5) is typically called “Online projected gradient descent”. It is straightforward to show that it has the same guarantees as Algorithm 3 from Theorem 5.

**Algorithm 5**

**Online projected gradient descent**

**input** Closed convex set $S \subseteq \mathbb{R}^d$, step size $\eta > 0$.

1: Pick $w_1 \in S$.
2: for $t = 1$ to $T$ do
3: Receive feedback: $f_t(w_t)$ and $\lambda_t \in \partial f_t(w_t)$ for some convex function $f_t: S \to \mathbb{R}$.
4: Incur loss: $f_t(w_t)$.
5: Update: $w_{t+1} := \arg\min_{w\in S} \|w - (w_t - \eta \lambda_t)\|^2$.
6: end for

4 Duality

In “Follow the regularized leader”, we made a seemingly arbitrary decision to use the squared Euclidean norm as a regularizer. But, in fact, this choice determined the qualities of the resulting regret bound in Theorems 4 and 5 (or more specifically, Corollaries 1 and 2): the bound depends on the Euclidean norms of the subgradients of the loss functions $f_1, f_2, \ldots, f_T$, as well as the Euclidean norm of the comparator $u$. What if the subgradients of the loss functions actually have large Euclidean norms, but are bounded in $l_\infty$ norm? What if we are only interested in comparators that in fact have small $l_1$ norm (in addition to having small Euclidean norm)? In such cases, we should use a different regularizer other than squared Euclidean norm.

We shall arrive at these other regularizers by studying the duality properties of convex functions.
4.1 Convex conjugates

Every function $f: \mathbb{R}^d \to \bar{\mathbb{R}}$ is associated with another function called its convex conjugate (or Fenchel conjugate) $f^*: \mathbb{R}^d \to \mathbb{R}$:

$$f^*(y) := \sup\left\{\langle y, x \rangle - f(x) : x \in \mathbb{R}^d\right\} = \sup\left\{\langle y, x \rangle - f(x) : x \in \text{dom } f\right\}.$$

(where $\text{dom } f := \{x \in \mathbb{R}^d : f(x) < \infty\}$ is the domain of $f$). Observe that $f^*$ is a supremum of linear functions of $y$, and therefore $f^*$ is convex. In fact, convex conjugates have the following properties.

**Theorem 6.** Assume $f: \mathbb{R}^d \to \bar{\mathbb{R}}$ is proper.

1. $f^*$ is closed and convex.
2. $f^{**} = f$ iff $f$ is closed and convex.

Henceforth, we again implicitly assume that the convex functions throughout Section 4 are closed and proper.

When $f$ is convex, one should think of the domain of $f^*$ as the space to which the subgradients of $f$ belong. Points from the domains of $f$ and $f^*$ interact in inner products, which can be understood through the Fenchel-Young inequality.

**Lemma 2** (Fenchel-Young). For any function $f: \mathbb{R}^d \to \bar{\mathbb{R}}$ and any $x, y \in \mathbb{R}^d$,

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

with equality if and only if $x \in \text{dom } f$ and $y \in \partial f(x)$.

**Proof.** The inequality follows immediately from the definition of convex conjugates. Moreover,

$$x \in \text{dom } f \land y \in \partial f(x)$$

$$\iff f(x') \geq f(x) + \langle y, x' - x \rangle \quad \forall x' \in \mathbb{R}^d$$

$$\iff \langle y, x \rangle - f(x) \geq \langle y, x' \rangle - f(x') \quad \forall x' \in \mathbb{R}^d$$

$$\iff \langle y, x \rangle - f(x) \geq \sup\left\{\langle y, x' \rangle - f(x') : x' \in \mathbb{R}^d\right\} = f^*(y).$$

Here are some examples of convex conjugate pairs. (Below, we use the convention $0 \ln 0 = 0$.)
1. \( f(x) = |x|^p/p \) for \( x \in \mathbb{R} \) (where \( 1 < p < \infty \)); \( f^*(y) = |y|^q/q \) for \( y \in \mathbb{R} \) (where \( 1/p + 1/q = 1 \)).

2. \( f(x) = -\ln(x) \) for \( x > 0 \); \( f^*(y) = -1 - \ln(-y) \) for \( x < 0 \).

3. \( f(x) = \ln(1 + \exp(x)) \) for \( x \in \mathbb{R} \); \( f^*(y) = y \ln(y) + (1 - y) \ln(1 - y) \) for \( y \in [0, 1] \).

4. \( f(x) = \exp(x) \) for \( x \in \mathbb{R} \); \( f^*(y) = y \ln(y) - y \) for \( y \geq 0 \).

5. \( f(x) = 0 \) for \( x \in \mathbb{R} \); \( f^*(y) = 0 \) for \( y \in \{0\} \).

6. \( f(x) = c \cdot g(ax + b) \) (where \( a \neq 0 \) and \( c > 0 \)); \( f^*(y) = c \cdot g^*(y/(ac)) - by \).

### 4.2 Norms

We now discuss norms on \( \mathbb{R}^d \). Recall that a norm on \( \mathbb{R}^d \) is a function \( \| \cdot \| : \mathbb{R}^d \to \mathbb{R} \) that satisfies, for all \( c \in \mathbb{R} \) and all \( u, v \in \mathbb{R}^d \), (i) \( \| cu \| = |c| \| u \| \), (ii) \( \| u + v \| \leq \| u \| + \| v \| \), and (iii) \( \| u \| = 0 \Rightarrow u = 0 \).

So far, we have used \( \| \cdot \| \) to denote the Euclidean norm, which is given by the square-root of the standard inner product between a vector \( x \in \mathbb{R}^d \) and itself: \( \| x \| := \sqrt{\langle x, x \rangle} \). But the Euclidean norm is just one a member of the infinite family of \( l_p \) norms \( 1 \leq p < \infty \):

\[
\| x \|_p := \left( \sum_{i=1}^{d} |x_i|^p \right)^{1/p}
\]

(for \( p = \infty \), we define \( \| x \|_\infty := \max_{i \in [d]} |x_i| \)); the Euclidean norm is the \( l_2 \) norm. Another infinite family of norms are the Mahalanobis norms: for any symmetric positive definite matrix \( A \in \mathbb{R}^{d \times d} \),

\[
\| X \|_A := \sqrt{\langle Ax, x \rangle};
\]

the Euclidean norm is obtained with \( A = I \).

Norms come in pairs. Let \( \| \cdot \| \) denote a generic norm on \( \mathbb{R}^d \). The associated dual norm, denoted by \( \| \cdot \|_* \), is given by

\[
\| y \|_* := \max \{ \langle y, x \rangle : \| x \| \leq 1 \}.
\]

It turns out the norm that is dual to \( \| \cdot \|_* \) is simply \( \| \cdot \| \). The \( l_p \) norm and the \( l_q \) norm, where \( 1/p + 1/q = 1 \), and dual to each other. The Mahalanobis norms given by the matrices \( A \) and \( A^{-1} \) are dual to each other. The Euclidean norm is dual to itself.
Norms on $\mathbb{R}^d$ are convex functions (because of the triangle inequality and homogeneity). Let $B_{\|\cdot\|} := \{y \in \mathbb{R}^d : \|y\| \leq 1\}$ is the unit ball for the dual norm $\|\cdot\|_*$. If $f(x) := \|x\|$ for some norm $\|\cdot\|$, then the subgradient of $f$ at $x$ is

$$\partial f(x) = \text{conv}\{y \in B_{\|\cdot\|_*} : \langle y, x \rangle = \|x\|\},$$

and the convex conjugate of $f$ is given by

$$f^*(y) = I_{B_{\|\cdot\|_*}}(y).$$

### 4.3 Strict and strong convexity

Recall that a function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is convex if and only if

$$f((1 - \alpha)x + \alpha x') \leq (1 - \alpha)f(x) + \alpha f(x') \quad \forall x, x' \in \text{dom } f, \alpha \in [0, 1].$$

(This is from Proposition 1, which is essentially Jensen’s inequality.) If the inequality above is always strict whenever $x \neq x'$, then we call the function strictly convex. Linear functions, for instance, are convex but not strictly convex; the function $x \mapsto \exp(x)$ is strictly convex. An equivalent criterion for strict convexity is that

$$f(x') \geq f(x) + \langle \lambda, x' - x \rangle, \quad \forall x, x' \in \text{dom } f, \lambda \in \partial f(x)$$

with strict inequality whenever $x \neq x'$. Thus, a strictly convex function lies strictly above its affine approximation at every $x \in \text{dom } f$ (well, it lies strictly above everywhere except at $x$).

Let $\sigma > 0$ and $\|\cdot\|$ be a norm on $\mathbb{R}^d$. We say a function $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ is $\sigma$-strongly convex with respect to a norm $\|\cdot\|$ if

$$f((1 - \alpha)x + \alpha x') \leq (1 - \alpha)f(x) + \alpha f(x') - \alpha(1 - \alpha)\frac{\sigma}{2}\|x - x'\|^2$$

$$\forall x, x' \in \text{dom } f, \alpha \in [0, 1].$$

The parameter $\sigma$ is called the modulus of strong convexity; it quantifies how uniformly non-linear the function is. An equivalent criterion for $\sigma$-strong convexity with respect to $\|\cdot\|$ is that

$$f(x') \geq f(x) + \langle \lambda, x' - x \rangle + \frac{\sigma}{2}\|x' - x\|^2, \quad \forall x, x' \in \text{dom } f, \lambda \in \partial f(x).$$

Thus, a $\sigma$-strongly convex function lies not only above its affine approximation at every $x \in \text{dom } f$, but in fact it lies above a quadratic approximation at every $x \in \text{dom } f$.

Some typical examples of strongly convex functions are as follows.
1. For $1 < p \leq 2$, $f(x) := 0.5\|x\|_p^2/(p - 1)$ with domain $\mathbb{R}^d$ is $1$-strongly convex with respect to the $l_p$ norm $\| \cdot \|_p$.

2. $f(x) := \sum_{i=1}^d x_i \ln x_i$ (negative entropy) with domain $\Delta^{d-1}$ (the probability simplex) is $1$-strongly convex with respect to the $l_1$ norm $\| \cdot \|_1$.

3. For a symmetric positive definite matrix $A \in \mathbb{R}^{d \times d}$, $f(x) := 0.5\langle Ax, x \rangle$ with domain $\mathbb{R}^d$ is $\sigma$-strongly convex with respect to the Euclidean norm $\| \cdot \|_2$, where $\sigma$ is the smallest eigenvalue of $A$. It is $1$-strongly convex with respect to the Mahalanobis norm $\| \cdot \|_A$.

Also, note that if $f$ is $1$-strongly convex with respect to a norm $\| \cdot \|$, then for any $\sigma > 0$, the function $g := \sigma f$ is $\sigma$-strongly convex with respect to the same norm $\| \cdot \|$. Because of this scaling property, we suffices to just understand $1$-strongly convex functions. We’ll simply say “strongly convex” when we mean “$1$-strongly convex”.

The first consequence of strong convexity is that it guarantees the existence of a unique minimizer.

**Proposition 5.** Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be strongly convex with respect to a norm $\| \cdot \|$. Then $f$ has a unique minimizer in $\text{dom } f$.

Strongly convex functions are precisely the functions that make for good regularizers in “Follow the regularized leader”. This is primarily due to the properties of their convex conjugates.

Before describing these properties, we first record the convex conjugates of the examples from above.

1. For $1 < p \leq 2$, $f(x) := 0.5\|x\|_p^2/(p - 1)$ with domain $\mathbb{R}^d$ has convex conjugate $f^*(y) := 0.5\|x\|_q^2/(q - 1)$ where $1/p + 1/q = 1$.

2. $f(x) := \sum_{i=1}^d x_i \ln x_i$ with domain $\Delta^{d-1}$ has convex conjugate $f^*(y) = \ln \sum_{i=1}^d \exp(y_i)$.

3. For a symmetric positive definite matrix $A \in \mathbb{R}^{d \times d}$, $f(x) := 0.5\langle Ax, x \rangle$ with domain $\mathbb{R}^d$ has convex conjugate $f^*(y) = 0.5\langle y, A^{-1}y \rangle$.

**Theorem 7.** Let $f : \mathbb{R}^d \to \bar{\mathbb{R}}$ be strongly convex with respect to a norm $\| \cdot \|$.

1. $f^*$ is differentiable on $\mathbb{R}^d$.

2. $\nabla f^*(y) = \arg \max_{x \in \mathbb{R}^d} \langle y, x \rangle - f(x)$.

3. $y \in \partial f(\nabla f^*(y))$ for all $y \in \mathbb{R}^d$. 

20
4. \[ \| \nabla f^*(y) - \nabla f^*(y') \| \leq \| y - y' \|, \text{ for all } y, y' \in \mathbb{R}^d. \]

5. \[ f^*(y') \leq f^*(y) + \langle y - y, \nabla f^*(y) \rangle + \| y' - y \|^2/2 \text{ for all } y, y' \in \mathbb{R}^d. \]

Proof. Since \( f \) is closed and strongly convex, for all \( y \in \mathbb{R}^d \), the function \( x \mapsto \langle y, x \rangle - f(x) \) has a unique maximizer (Proposition 5)—call it \( v(y) \). By definition of \( f^* \), for any \( y \in \mathbb{R}^d \),

\[ f^*(y') = \sup_{x \in \mathbb{R}^d} \{ \langle y', x \rangle - f(x) \} \geq \langle y', v(y) \rangle - f(v(y)) \quad \text{for all } y' \in \mathbb{R}^d. \]

Above, equality holds if \( y = y' \), in which case we have \( f^*(y) = \langle y, v(y) \rangle - f(v(y)) \). Therefore, the above inequality reads

\[
f^*(y') \geq \langle y', v(y) \rangle - \langle (y, v(y)) - f^*(y) \rangle = f^*(y) + \langle y' - y, v(y) \rangle \quad \text{for all } y' \in \mathbb{R}^d.
\]

This means that \( v(y) \in \partial f^*(y) \). Now consider any other \( v \in \partial f^*(y) \). By Fenchel-Young (Lemma 2),

\[ f^*(y) = \langle y, v(y) \rangle - f(v(y)) = \langle y, v \rangle - f(v). \]

The uniqueness of the maximizer of \( x \mapsto \langle y, x \rangle - f(x) \) (established above) implies that \( v = v(y) \). Therefore \( \partial f^*(y) = \{ v(y) \} \) for every \( y \in \mathbb{R}^d \), so \( f^* \) is differentiable and \( \nabla f^*(y) = v(y) = \arg \max_{x \in \mathbb{R}^d} \langle y, x \rangle - f(x) \). Thus the first two claims are proved.

For the third claim, observe that for all \( x \in \text{dom } f \) and \( y \in \mathbb{R}^d \), the optimality of \( v(y) \) implies that

\[
f(x) - f(v(y)) = f(x) + \langle y, v(y) \rangle - f(v(y)) - \langle y, v(y) \rangle \geq f(x) + \langle y, x \rangle - f(x) - \langle y, v(y) \rangle = \langle y, x - v(y) \rangle.
\]

This implies that \( y \in \partial f(v(y)) \) as required.

For the fourth claim, let \( x := \nabla f^*(y) \) and \( x' := \nabla f^*(y') \), so \( y \in \partial f(x) \) and \( y' \in \partial f(x') \) by the third claim. Define \( x_\alpha := (1 - \alpha)x + \alpha x' \) for \( \alpha \in (0, 1) \). Since \( f \) is strongly convex with respect to \( \| \cdot \| \),

\[
f(x_\alpha) \geq f(x) + \langle y, x_\alpha - x \rangle + \frac{1}{2} \| x_\alpha - x \|^2
\]

\[
= f(x) - \alpha \langle y, x - x' \rangle + \alpha^2 \frac{1}{2} \| x - x' \|^2
\]

\[
f(x_\alpha) \geq f(x') + \langle y', x_\alpha - x' \rangle + \frac{1}{2} \| x_\alpha - x' \|^2
\]

\[
= f(x') + (1 - \alpha) \langle y', x - x' \rangle + (1 - \alpha)^2 \frac{1}{2} \| x - x' \|^2.
\]

21
Multiplying both sides of these two inequalities by $1 - \alpha$ and $\alpha$, respectively, and adding them together gives

$$f(x_\alpha) \geq (1 - \alpha)f(x) + \alpha f(x') - \alpha(1 - \alpha)\langle y - y', x - x' \rangle$$

$$+ \alpha(1 - \alpha)\frac{1}{2}\|x - x'\|^2.$$ 

Also by strong convexity of $f$,

$$f(x_\alpha) \leq (1 - \alpha)f(x) + \alpha f(x') - \alpha(1 - \alpha)\frac{1}{2}\|x - x'\|^2.$$ 

Combining these last two inequalities, rearranging, and canceling $\alpha(1 - \alpha)$ gives

$$\langle y - y', x - x' \rangle \geq \|x - x'\|^2$$ 

By definition of dual norm $\|\cdot\|_*$,

$$\langle y - y', x - x' \rangle \leq \|y - y'\|_*\|x - x'\|.$$ 

Combining these last two inequalities and simplifying gives

$$\|y - y'\|_* \geq \|x - x'\|.$$ 

For the final claim, let $\delta := y' - y$, and observe that by the fundamental theorem of calculus,

$$f^*(y + \delta) - f^*(y) - \langle \delta, \nabla f^*(y) \rangle = \int_0^1 \langle \delta, \nabla f^*(y + t\delta) - \nabla f^*(y) \rangle dt.$$ 

By the definition of the dual norm and the fourth claim, for any $t \geq 0$,

$$\langle \delta, \nabla f^*(y + t\delta) - \nabla f^*(y) \rangle \leq \|\delta\|_*\|\nabla f^*(y + t\delta) - \nabla f^*(y)\| \leq t\|\delta\|_*^2,$$

so

$$\int_0^1 \langle \delta, \nabla f^*(y + t\delta) - \nabla f^*(y) \rangle dt \leq \frac{1}{2}\|\delta\|_*^2$$

as required. 

Theorem 7 quite explicitly reveals the duality of $f$ and $f^*$. Suppose $f$ is differentiable. Then the third claim in Theorem 7 says $y = \nabla f(\nabla f^*(y))$ for all $y \in \mathbb{R}^d$. If, in addition, $\{\nabla f(x) : x \in \text{dom} f\} = \mathbb{R}^d$, then it turns out that the maps $x \mapsto \nabla f(x)$ and $y \mapsto \nabla f^*(y)$ are inverses of each other. In this case, the domains of $f$ and $f^*$ can be thought of as two different spaces which are connected by the map $\nabla f$ and its inverse $\nabla f^*$. 

22
The fifth claim in Theorem 7 looks like the definition of strong convexity, except the inequality is reversed: it says that $f^*$ is bounded above by a quadratic approximation at any point $y \in \mathbb{R}^d$. This property is called smoothness—we say that $f^*$ is smooth with respect to the norm $\| \cdot \|_*$.

We can also quantify smoothness in the same way we quantify strong convexity. Recall that for any $\sigma > 0$, the convex conjugate of $f(x) = \sigma g(x)$ is $f^*(y) = \sigma g^*(y/\sigma)$. Let us assume $g$ is 1-strongly convex with respect to $\| \cdot \|$. Then $f$ is $\sigma$-strongly convex with respect to $\| \cdot \|$. Moreover, for any $y, y' \in \mathbb{R}^d$,

$$f^*(y') = \sigma g^*(y'/\sigma) \leq \sigma \left( g^*(y/\sigma) + \langle (y' - y)/\sigma, \nabla g^*(y/\sigma) \rangle + \frac{1}{2\sigma} \| (y' - y)/\sigma \|^2_* \right) = f^*(y) + \langle y' - y, \nabla f^*(y) \rangle + \frac{1}{2\sigma} \| y' - y \|^2_*.$$

The second step uses the smoothness property of $g$, and the last step uses the fact that $\nabla f^*(y) = \nabla g^*(y/\sigma)$ by the chain rule. We call a function $f^*$ satisfying the above inequality $(1/\sigma)$-smooth with respect to the norm $\| \cdot \|_*$. In fact, it can be shown that if $f^*$ is $(1/\sigma)$-smooth with respect to $\| \cdot \|_*$, then $f$ must be $\sigma$-strongly convex with respect to $\| \cdot \|$. Hence “strong convexity” and “smoothness” are dual to each other in this sense.

5 Primal-dual algorithms

We now study online learning algorithms that take advantage of other convex regularization functions and their duality properties.

5.1 Online dual averaging and mirror maps

We first consider a general algorithm called “Online dual averaging” (Algorithm 6), which is precisely the same as “Follow the regularized leader” with a general strongly convex regularization function $\phi : \mathbb{R}^d \to \mathbb{R}$. Note that we can incorporate the choice of a convex domain $S$ different from $\mathbb{R}^d$ into the regularization function $\phi$ by setting $\phi(w) := +\infty$ for all $w \notin S$. (We will see the algorithmic consequences of this later.)
Algorithm 6 Online dual averaging

\textbf{input} Regularization parameter $\eta > 0$, regularization function $\phi : \mathbb{R}^d \to \mathbb{R}$.

1: Let $w_1 := \arg \min_{w \in \mathbb{R}^d} \phi(w)$.
2: for $t = 1$ to $T$
3: \hspace{1em} Receive feedback: $f_t(w_t)$ and $\lambda_t \in \partial f_t(w_t)$ for some convex function $f_t : \mathbb{R}^d \to \mathbb{R}$.
4: \hspace{1em} Incur loss: $f_t(w_t)$.
5: \hspace{1em} Update:
6: \hspace{2em} $w_{t+1} := \arg \min_{w \in \mathbb{R}^d} \eta \sum_{i=1}^{t} \langle \lambda_i, w \rangle + \phi(w)$. (7)
7: \hspace{1em} end for

When $\phi(w) = 0.5 \|w\|_2^2$, the update (7) was simply determined by minimizing a convex quadratic function, and hence it could be written down in closed form. What about for a general strongly convex $\phi$? Observe that

\[
\min_{w \in \mathbb{R}^d} \left\{ \eta \sum_{i=1}^{t} \langle \lambda_i, w \rangle + \phi(w) \right\} = - \max_{w \in \mathbb{R}^d} \left\{ -\eta \sum_{i=1}^{t} \langle \lambda_i, w \rangle - \phi(w) \right\}
\]

\[
= - \max_{w \in \mathbb{R}^d} \left\{ -\eta \sum_{i=1}^{t} \lambda_i, w \right\} - \phi(w) \}
\]

\[
= -\phi^* \left( -\eta \sum_{i=1}^{t} \lambda_i \right).
\]

And which $w \in \mathbb{R}^d$ achieves the maximum in the last step? Theorem 7 tells us that it is precisely given by the \textit{mirror map} $\nabla \phi^*$:

\[
w_{t+1} := \nabla \phi^* \left( -\eta \sum_{i=1}^{t} \lambda_i \right). \quad (8)
\]

\textbf{Example 2} (Squared $l_p$ norm regularization). Fix $1 < p \leq 2$. Let $\phi(w) := 0.5 \|w\|_p^2/(p-1)$ for all $w \in \mathbb{R}^d$. Then $\phi$ is strongly convex on $\mathbb{R}^d$ with respect to the $l_p$ norm $\| \cdot \|_p$. The convex conjugate of $\phi$ is $\phi^*(m) = 0.5 \|w\|_q^2/(q-1)$ where $1/p + 1/q = 1$ (note that $q \geq 2$ since $p \leq 2$), and its gradient is

\[
\nabla \phi^*(m)_j = \text{sign}(m[j]) \frac{|m[j]|^{q-1}}{\|m\|_q^{q-2}} \text{ for all } j \in [d].
\]

Therefore, to compute the update (8), we may accumulate the subgradients

\[
m_{t+1} := m_t - \eta \lambda_t
\]

24
and then set $w_{t+1}$ using

$$w_{t+1}[j] := \text{sign}(m_{t+1}[j]) \frac{|m_{t+1}[j]|^{q-1}}{\|m_{t+1}\|_q^q} \quad \text{for all } j \in [d].$$

**Example 3** (Negative entropy regularization). Let $\phi(w) := \sum_{j=1}^d w[j] \ln w[j]$ for $w = (w[1], w[2], \ldots, w[d]) \in \Delta^{d-1}$, and $\phi(w) := +\infty$ for $w \notin \Delta^{d-1}$. Then $\phi$ is strongly convex on $\Delta^{d-1}$ with respect to the $l_1$ norm $\|\cdot\|_1$. The convex conjugate of $\phi$ is $\phi^*(m) = \ln \sum_{j=1}^d \exp(m[j])$ for $m = (m[1], m[2], \ldots, m[d]) \in \mathbb{R}^d$, and its gradient is

$$\nabla \phi^*(m)_j = \frac{\exp(m[j])}{\sum_{k=1}^d \exp(m[k])} \quad \text{for all } j \in [d].$$

Therefore, the update (8) can be written as

$$w_{t+1}[j] := \frac{\exp(-\eta \sum_{i=1}^t \lambda_i[j])}{\sum_{k=1}^d \exp(-\eta \sum_{i=1}^t \lambda_i[k])} \quad \text{for all } j \in [d].$$

It is easy to see that $w_{t+1} \in \Delta^{d-1}$. Conveniently,

$$w_{t+1}[j] \propto w_t[j] \exp(-\eta \lambda_t[j]).$$

**Example 4** (Squared $l_2$ norm regularization on a general convex domain). Let $\phi(w) := 0.5\|w\|_2^2 + I_S(w)$ for some convex $S \subseteq \mathbb{R}^d$, which is strongly convex on $S$ with respect to the Euclidean norm. Its convex conjugate is

$$\phi^*(m) = \max_{w \in \mathbb{R}^d} \langle m, w \rangle - \frac{1}{2} \|w\|_2^2 - I_S(w)$$

$$= -\min_{w \in S} (-m, w) + \frac{1}{2} \|w\|_2^2$$

$$= -\min_{w \in S} \left( \frac{1}{2} \|w - m\|_2^2 - \frac{1}{2} \|m\|_2^2 \right).$$

The minimum in the last step is achieved at $w := \arg \min_{w \in S} \|w' - m\|_2 = \Pi_S(m)$, i.e., the Euclidean projection of $m$ onto $S$. Therefore

$$\nabla \phi^*(m) = \Pi_S(m),$$

and the update (8) is

$$w_{t+1} := \Pi_S \left( -\eta \sum_{i=1}^t \lambda_i \right).$$
Theorem 8. Let $\phi: \mathbb{R}^d \to \mathbb{R}$ be strongly convex on its domain with respect to a norm $\| \cdot \|$. Let $f_1, f_2, \ldots, f_T: \mathbb{R}^d \to \mathbb{R}$ be any sequence of convex loss functions. The total loss incurred by a learner using Algorithm 6 ("Online dual averaging") with regularization parameter $\eta > 0$ is bounded as

$$\sum_{t=1}^{T} f_t(w_t) \leq \sum_{t=1}^{T} f_t(u) + \frac{\phi(u) - \phi(w_1)}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\lambda_t\|^2_* \quad \text{for all } u \in \text{dom } \phi.$$  

Proof. Let $m_t := -\eta \sum_{i=1}^{t-1} \lambda_i$, so $w_t = \arg\max_{w \in \mathbb{R}^d} \langle m_t, w \rangle - \phi(w)$. Since $\phi$ is strongly convex, Theorem 7 implies that $w_t = \nabla \phi^*(m_t)$. Also by Theorem 7, $\phi^*$ is smooth with respect to $\| \cdot \|_*$. Hence

$$\phi^*(m_{t+1}) - \phi^*(m_t) \leq \langle m_{t+1} - m_t, \nabla \phi^*(m_t) \rangle + \frac{1}{2} \| m_{t+1} - m_t \|^2_*$$

$$= -\eta \langle \lambda_t, w_t \rangle + \frac{\eta^2}{2} \| \lambda_t \|^2_*.$$  

Therefore, since $m_1 = 0$,

$$\eta \sum_{t=1}^{T} \langle \lambda_t, w_t \rangle \leq \phi^*(0) - \phi^*(m_{T+1}) + \frac{\eta^2}{2} \sum_{t=1}^{T} \| \lambda_t \|^2_*.$$  

By the Fenchel-Young inequality (Lemma 2), for any $u \in \text{dom } \phi$,

$$\phi^*(m_{T+1}) \geq -\phi(u) + \langle m_{T+1}, u \rangle = -\phi(u) - \eta \sum_{t=1}^{T} \langle \lambda_t, u \rangle.$$  

Therefore

$$\eta \sum_{t=1}^{T} \langle \lambda_t, w_t - u \rangle \leq \phi^*(0) + \phi(u) + \frac{\eta^2}{2} \sum_{t=1}^{T} \| \lambda_t \|^2_*.$$  

To complete the proof, we simply use the fact $f_t(w_t) - f_t(u) \leq \langle \lambda_t, w_t - u \rangle$ (since $\lambda_t \in \partial f_t(w_t)$), as well as the fact that $\phi^*(0) = -\phi(w_1)$ (from the definition of $w_1$).

We leave it as a simple exercise to derive sublinear regret bounds from Theorem 8.
5.2 Bregman divergences

A continuous function \( \phi : \mathbb{R}^d \to \mathbb{R} \) is Legendre if

- \( \phi \) is strictly convex and differentiable on its domain; and
- for any sequence \((x_t)_{t \in \mathbb{N}}\) with \( \lim x_t \in \text{bd dom } \phi \), \( \lim \| \nabla \phi(x_t) \| = +\infty \).

Note that negative entropy on the probability simplex (Example 3) is not Legendre, although generalized negative entropy \( \phi(x) := \sum_{i=1}^d x_i \ln x_i - \sum_{i=1}^d x_i \) over \( \mathbb{R}_+^d \) is. (Recall that we assume the domain of \( \phi \) is closed.)

There is a general class of divergence measures that generalize squared Euclidean distance and relative entropy divergence, generated by Legendre functions: the Bregman divergences. The Bregman divergence associated with a Legendre function \( \phi : \mathbb{R}^d \to \mathbb{R} \) is given by

\[
D_{\phi}(x', x) := \phi(x') - \phi(x) - \langle \nabla \phi(x), x' - x \rangle
\]

for all \( x' \in \text{dom } \phi \) and \( x \in \text{int dom } \phi \). This is simply the difference between \( \phi(x') \) and the first-order approximation of \( \phi(x') \) around \( x \). Since \( \phi \) is strictly convex, \( D_{\phi}(x', x) \geq 0 \) with equality if and only if \( x' = x \). Note that if \( \phi \) is strongly convex with respect to a norm \( \| \cdot \| \), then

\[
D_{\phi}(x', x) \geq \frac{1}{2} \| x' - x \|^2.
\]

We can generalize the Euclidean projection onto a closed convex set \( S \subseteq \text{dom } \phi \) to the Bregman projection onto \( S \).

**Proposition 6.** Let \( \phi \) be Legendre and \( S \subseteq \text{int dom } \phi \) be a closed convex set. For any \( x \in \text{int dom } \phi \), the minimizer of \( x' \mapsto D_{\phi}(x', x) \) over \( x' \in S \) exists and is unique: i.e.,

\[
\Pi_{\phi,S}(x) := \arg \min_{x' \in S} D_{\phi}(x', x)
\]

is well-defined.

Bregman divergences and projections satisfy the following kind of Pythagorean relation that is well-known in the case of squared Euclidean distance.

**Lemma 3.** Let \( \phi \) be Legendre, and \( S \subseteq \text{int dom } \phi \) be a closed convex set. For any \( x \in \text{int dom } \phi \) and \( x' \in S \),

\[
D_{\phi}(x', \Pi_{\phi,S}(x)) + D_{\phi}(\Pi_{\phi,S}(x), x) \leq D_{\phi}(x', x).
\]
Proof. The gradient of \( f : x' \mapsto D_\phi(x', x) \) with respect to \( x' \) is
\[
\nabla f(x') = \nabla \phi(x') - \nabla \phi(x).
\]
By the optimality condition from Proposition 4, \( x_* := \Pi_{\phi,S}(x) \) satisfies
\[
\langle \nabla f(x_*), x_* - x' \rangle = \langle \nabla \phi(x_*), x_* - x' \rangle \leq 0 \quad \forall x \in S.
\]
Therefore, for any \( x \in S \),
\[
D_\phi(x', x) + D_\phi(x, x) = \phi(x') - \phi(x) - \langle \nabla \phi(x_*), x' - x_* \rangle
+ \phi(x_*) - \phi(x) - \langle \nabla \phi(x), x_* - x \rangle
= D_\phi(x', x) + \langle \nabla \phi(x_*) - \nabla \phi(x), x_* - x' \rangle
\leq D_\phi(x', x).
\]

5.3 Online mirror descent

“Online mirror descent” (Algorithm 7) is another primal-dual algorithm, much like “Online dual averaging” (Algorithm 6).

**Algorithm 7 Online mirror descent**

**input** Legendre regularization function \( \phi : \mathbb{R}^d \to \bar{\mathbb{R}} \), closed convex set \( S \subseteq \text{int dom } \phi \), step size \( \eta > 0 \).

1: Let \( w_1 := \arg\min_{w \in S} \phi(w) \).
2: for \( t = 1 \) to \( T \) do
3: Receive feedback: \( f_t(w_t) \) and \( \lambda_t \in \partial f_t(w_t) \) for some convex function \( f_t : \mathbb{R}^d \to \mathbb{R} \).
4: Incur loss: \( f_t(w_t) \).
5: Update:
   \[
   w_{t+1} := \arg\min_{w \in S} \eta \langle \lambda_t, w \rangle + D_\phi(w, w_t). \tag{9}
   \]
6: end for

The update step in “Online mirror descent” is usually written as
\[
\nabla \phi(\tilde{w}_{t+1}) := \nabla \phi(w_t) - \eta \lambda_t;
\]
\[
w_{t+1} := \Pi_{\phi,S}(\tilde{w}_{t+1}). \tag{10}
\]

The first step is a gradient step in the dual space to get \( \nabla \phi(\tilde{w}_{t+1}) \); then its primal counterpart \( \tilde{w}_{t+1} \) is projected onto \( S \) using the Bregman projection...
to get $w_{t+1}$. To see why (10) is equivalent to (9), observe that

$$w_{t+1} = \arg \min_{w \in S} \eta \langle \lambda_t, w \rangle + D_\phi(w, w_t)$$

$$= \arg \min_{w \in S} \eta \langle \lambda_t, w \rangle + (\phi(w) - \phi(w_t) - \langle \nabla \phi(w_t), w - w_t \rangle)$$

$$= \arg \min_{w \in S} \phi(w) - \langle \nabla \phi(w_t), w \rangle$$

$$= \arg \min_{w \in S} \phi(w) - \langle \nabla \phi(w_{t+1}), w \rangle$$

$$= \arg \min_{w \in S} \phi(w) - \phi(w_{t+1}) - \langle \nabla \phi(w_{t+1}), w - w_{t+1} \rangle$$

$$= \arg \min_{w \in S} D_\phi(w, \tilde{w}_{t+1})$$

$$= \pi_{\phi,S}(\tilde{w}_{t+1}).$$

**Theorem 9.** Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be Legendre and strongly convex with respect to a norm $\| \cdot \|$ and $S \subseteq \text{int dom } \phi$ be a closed convex set. Let $f_1, f_2, \ldots, f_T : \mathbb{R}^d \to \mathbb{R}$ be any sequence of convex loss functions. The total loss incurred by a learner using Algorithm 4 (“Online mirror descent”) with step size $\eta > 0$ is bounded as

$$\sum_{t=1}^T f_t(w_t) \leq \sum_{t=1}^T f_t(u) + \frac{\eta}{2} \sum_{t=1}^T \| \lambda_t \|^2 + \frac{D_\phi(u, w_1) - D_\phi(u, w_{T+1})}{\eta}$$

for all $u \in S$.

**Proof.** Fix any $u \in S$. Consider the following decomposition of $\langle \lambda_t, w_t - u \rangle$:

$$\langle \lambda_t, w_t - u \rangle = \langle \lambda_t, w_t - \tilde{w}_{t+1} \rangle + \langle \lambda_t, \tilde{w}_{t+1} - u \rangle.$$

We bound the first term as follows:

$$\langle \lambda_t, w_t - \tilde{w}_{t+1} \rangle \leq \| \lambda_t \|_* \| w_t - \tilde{w}_{t+1} \|$$

$$\leq \frac{\eta}{2} \| \lambda_t \|^2 + \frac{1}{2\eta} \| w_t - \tilde{w}_{t+1} \|^2$$

$$\leq \frac{\eta}{2} \| \lambda_t \|^2 + \frac{1}{\eta} D_\phi(\tilde{w}_{t+1}, w_t).$$

The first inequality follows by the definition of dual norm, the second inequality is by AM/GM, and the third inequality follows by strong convexity of $\phi$ with respect to the norm $\| \cdot \|$.

29
To bound the second term, we shall use the fact that

$$\lambda_t = \frac{1}{\eta} (\nabla \phi(w_t) - \nabla \phi(\tilde{w}_{t+1}))$$

as seen from the update (10), so that

$$\langle \lambda_t, \tilde{w}_{t+1} - u \rangle = \frac{1}{\eta} (\nabla \phi(w_t) - \nabla \phi(\tilde{w}_{t+1}), \tilde{w}_{t+1} - u)$$

$$= \frac{1}{\eta} (D_\phi(u, w_t) - D_\phi(u, \tilde{w}_{t+1}) - D_\phi(\tilde{w}_{t+1}, w_t))$$

$$\leq \frac{1}{\eta} (D_\phi(u, w_t) - D_\phi(u, w_{t+1}) - D_\phi(\tilde{w}_{t+1}, w_t)). \quad (12)$$

The inequality follows from the fact that $u \in S$ and that $w_{t+1}$ is the Bregman projection of $\tilde{w}_{t+1}$ onto $S$.

Combining (11) and (12), we have

$$\langle \lambda_t, w_t - u \rangle \leq \eta \| \lambda_t \|^2 + \frac{1}{\eta} (D_\phi(u, w_t) - D_\phi(u, w_{t+1})).$$

Summing this up over $t$, and using the fact $f_t(w_t) - f_t(u) \leq \langle \lambda_t, w_t - u \rangle$ (by convexity of $f_t$) gives the claimed bound.

$$\square$$

### 6 Special loss functions

Often when one has more information about the loss functions $f_t$ (beyond convexity), it is possible to design better online convex optimization algorithms.

#### 6.1 exp-concave loss functions

One important property of convex loss functions is strong convexity. However, strong convexity is often too strong of a property to expect from loss functions (except when explicit statistical regularization is used). A weaker property is exp-concavity. We say a function $f: S \to \mathbb{R}$ is $\alpha$-exp-concave (for some $\alpha > 0$) if the function $x \mapsto \exp(-\alpha f(x))$ is a concave on $\text{dom } f$.

(Recall that a function is concave if its negation is convex.) For example, if $S = \Delta^{d-1}$, then for any $a \in \mathbb{R}^{d}_{++}$, the function $f(x) := -\ln(\langle a, x \rangle)$ is $1$-exp-concave on $S$.

Functions that are exp-concave can often be bounded below by a quadratic approximation in a weaker sense than strongly convex functions.
Lemma 4. Suppose $f$ is $\alpha$-exp-concave, and that $\beta \in (0, \alpha/2)$ satisfies

$$\langle \lambda, x - x' \rangle \leq \frac{1}{2\beta} \quad \forall x, x' \in \text{dom } f, \lambda \in \partial f(x).$$

Then for all $x, x' \in \text{dom } f$ and $\lambda \in \partial f(x)$,

$$f(x') \geq f(x) + \langle \lambda, x' - x \rangle + \frac{\beta}{2} \langle \lambda, x' - x \rangle^2.$$

Proof. Let $h(x) := \exp(-2\beta f(x))$, which is concave since $x \mapsto \exp(-\alpha f(x))$ is concave and $z \mapsto z^{2\beta/\alpha}$ is concave and non-decreasing on $\mathbb{R}_+$. Pick any $x, x' \in \text{dom } f$, and $\lambda \in \partial f(x)$. Observe that $\partial h(x) = \{-2\beta \exp(-2\beta f(x))\lambda : \lambda \in \partial f(x)\}$. Therefore, $g := -2\beta \exp(-2\beta f(x))\lambda \in \partial h(x)$, and

$$h(x') \leq h(x) + \langle g, x' - x \rangle.$$

Making substitutions, this reads

$$\exp(-2\beta f(x')) \leq \exp(-2\beta f(x)) - 2\beta \exp(-2\beta f(x))\langle \lambda, x' - x \rangle = \exp(-2\beta f(x))(1 - 2\beta \langle \lambda, x' - x \rangle).$$

Taking logarithm of both sides and dividing by $-2\beta$ gives

$$f(x') \geq f(x) - \frac{1}{2\beta} \ln(1 - 2\beta \langle \lambda, x' - x \rangle).$$

Recall that $\ln(1 + z) \leq z - z^2/4$ when $z \leq 1$. Therefore

$$f(x') \geq f(x) - \frac{1}{2\beta} (-2\beta \langle \lambda, x' - x \rangle - \beta^2 \langle \lambda, x' - x \rangle^2) = f(x) + \langle \lambda, x' - x \rangle + \frac{\beta}{2} \langle \lambda, x' - x \rangle^2. \quad \Box$$

In order to take advantage of exp-concavity, we may use a variant of “Online mirror descent” (Algorithm 7) where the regularization function changes from round to round. This algorithm is called “Online Newton step” (Algorithm 8), since it is reminiscent of Newton’s method (even though there are no Hessians involved).
Algorithm 8 Online Newton step

input Close convex set $S \subseteq \mathbb{R}^d$, step size $\eta > 0$, parameter $\varepsilon > 0$.
1: Let $A_1 := \varepsilon I$ and $w_1 := \arg \min_{w \in S} \|w\|^2_{A_1}$
2: for $t = 1$ to $T$ do
3: Receive feedback: $f_t(w_t)$ and $\lambda_t \in \partial f_t(w_t)$ for some convex function $f_t: S \to \mathbb{R}$.
4: Incur loss: $f_t(w_t)$.
5: Update: let $A_{t+1} := A_t + \lambda_t \lambda_t^T$ and
$$w_{t+1} := \arg \min_{w \in S} \eta \langle \lambda_t, w \rangle + \frac{1}{2} \|w - w_t\|^2_{A_{t+1}}.$$ (13)
6: end for

Define $\phi_t(w) := 0.5\|w\|^2_{A_t}$ (half of the squared Mahalanobis norm of $w$ with respect to $A_t$; it is easy to see that $A_t$ is symmetric positive definite). This function is Legendre and strongly convex with respect to $\|\cdot\|_{A_t}$. So the update (13) in Algorithm 8 (to obtain $w_{t+1}$) is the same as (10) except with $\phi$ replaced by $\phi_{t+1}$:
$$\nabla \phi_{t+1}(\tilde{w}_{t+1}) := \nabla \phi_{t+1}(w_t) - \eta \lambda_t;$$
$$w_{t+1} := \Pi_{\phi_{t+1},S}(\tilde{w}_{t+1}).$$ (14)
Since $\nabla \phi_{t+1}(w) = A_{t+1}w$, the update simplifies to
$$w_{t+1} := \arg \min_{w \in S} \|w - (w_t - \eta A_{t+1}^{-1} \lambda_t)\|^2_{A_{t+1}}.$$

Theorem 10. Let $f_1, f_2, \ldots, f_T: \mathbb{R}^d \to \mathbb{R}$ be any sequence of $\alpha$-exp-concave loss functions, and $S \subseteq \mathbb{R}^d$ be closed and convex. Assume $\beta \in (0, \alpha/2)$ satisfies $\langle \lambda, w - w' \rangle \leq 0.5/\beta$ for all $w, w' \in S$, all $t \in [T]$, and all $\lambda \in \partial f_t(w)$. Furthermore, assume that $\eta \geq 2/\beta$. The total loss incurred by a learner using Algorithm 8 (“Online Newton step”) with step size $\eta > 0$ and parameter $\varepsilon > 0$ is bounded as
$$\sum_{t=1}^T f_t(w_t) \leq \sum_{t=1}^T f_t(u) + \frac{\eta}{2} \frac{d}{2} \ln \left(1 + \frac{1}{\varepsilon} \sum_{t=1}^T \|\lambda_t\|^2_2\right) + \frac{\varepsilon}{\eta} \|u - w_1\|^2_2$$
for all $u \in S$.

Proof. The analysis of Algorithm 8 is very similar to that of Algorithm 7 except for two points. First, we must modify the analysis to handle the
time-dependent regularization function. Second, we must exploit the exp-
concavity of the loss functions \( f_t \). This latter point is handled using Lemma 4:

\[
f_t(w_t) - f_t(u) \leq \langle \lambda_t, w_t - u \rangle - \frac{\beta}{2} (w_t - u)^\top \lambda_t (w_t - u)
\]

(15)

(\text{where we use the shorthand } D_t := D_{\phi_t}).

Now we proceed with the analysis of \( \langle \lambda_t, w_t - u \rangle \) as before. Consider the following decomposition of \( \langle \lambda_t, w_t - u \rangle \):

\[
\langle \lambda_t, w_t - u \rangle = \langle \lambda_t, w_t - \tilde{w}_{t+1} \rangle + \langle \lambda_t, \tilde{w}_{t+1} - u \rangle.
\]

We bound the first term as follows:

\[
\langle \lambda_t, w_t - \tilde{w}_{t+1} \rangle \leq \| \lambda_t \|_{A_{t+1}} \| w_t - \tilde{w}_{t+1} \|_{A_{t+1}}
\]

\[
\leq \eta \| \lambda_t \|_{A_{t+1}}^2 + \frac{1}{2\eta} \| w_t - \tilde{w}_{t+1} \|_{A_{t+1}}^2
\]

\[
= \eta \| \lambda_t \|_{A_{t+1}}^2 + \frac{1}{\eta} D_{t+1}(\tilde{w}_{t+1}, w_t). \quad (16)
\]

The first inequality follows by the definition of dual norm, the second inequality is by AM/GM, and the third inequality follows by definition of \( D_{t+1} \).

To bound the second term, we shall use the fact that

\[
\lambda_t = \frac{1}{\eta} (\nabla \phi_{t+1}(w_t) - \nabla \phi_{t+1}(\tilde{w}_{t+1}))
\]
as seen from the update (14), so that
\[
\langle \lambda_t, \tilde{w}_{t+1} - u \rangle = \frac{1}{\eta} \langle \nabla \phi_{t+1}(w_t) - \nabla \phi_{t+1}(\tilde{w}_{t+1}), \tilde{w}_{t+1} - u \rangle \\
= \frac{1}{\eta} (D_{t+1}(u, w_t) - D_{t+1}(u, \tilde{w}_{t+1}) - D_{t+1}(\tilde{w}_{t+1}, w_t)) \\
\leq \frac{1}{\eta} (D_{t+1}(u, w_t) - D_{t+1}(u, w_{t+1}) - D_{t+1}(\tilde{w}_{t+1}, w_t)) \\
= \frac{1}{\eta} (D_t(u, w_t) - D_{t+1}(u, w_{t+1}) - D_{t+1}(\tilde{w}_{t+1}, w_t)) \\
+ \frac{1}{\eta} (D_{t+1}(u, w_t) - D_t(u, w_t)). \tag{17}
\]

The inequality follows from the fact that \( u \in S \) and that \( w_{t+1} \) is the Bregman projection of \( \tilde{w}_{t+1} \) onto \( S \).

Combining (16) and (17), we have
\[
\langle \lambda_t, w_t - u \rangle \leq \frac{\eta}{2} \|\lambda_t\|_{A_{t+1}^{-1}}^2 + \frac{1}{\eta} (D_t(u, w_t) - D_{t+1}(u, w_{t+1})) \\
+ \frac{1}{\eta} (D_{t+1}(u, w_t) - D_t(u, w_t)).
\]

Summing this over \( t \), and using (15), we have
\[
\sum_{t=1}^{T} f_t(w_t) - f_t(u) \leq \frac{\eta}{2} \sum_{t=1}^{T} \|\lambda_t\|_{A_{t+1}^{-1}}^2 + \frac{1}{\eta} (D_t(u, w_1) - D_{T+1}(u, w_{T+1})).
\]

It remains to bound the summation on the right-hand side. For this, we exploit a linear algebraic inequality
\[
\operatorname{tr}(X^{-1}(X - Y)) \leq \ln \frac{\det(X)}{\det(Y)}
\]
for positive definite matrices \( X \succeq Y > 0 \). Here, we use \( X := A_{t+1} \) and \( Y := A_t \), since
\[
\|\lambda_t\|_{A_{t+1}^{-1}}^2 = \lambda_t^\top A_{t+1}^{-1} \lambda_t \\
= \operatorname{tr}(A_{t+1}^{-1} \lambda_t \lambda_t^\top) \\
= \operatorname{tr}(A_{t+1}^{-1} (A_{t+1} - A_t)).
\]

34
Hence

\[
\sum_{t=1}^{T} \|\lambda_t\|^2_{A_{t+1}^{-1}} \leq \sum_{t=1}^{T} \ln \frac{\det(A_{t+1})}{\det(A_t)} = \ln \frac{\det(A_{T+1})}{\det(A_1)}
\]

\[
= \ln \frac{\det(\varepsilon I + \sum_{t=1}^{T} \lambda_t \lambda_t^T)}{\det(\varepsilon I)}
\]

\[
\leq \ln \left( \frac{\varepsilon + \sum_{t=1}^{T} \|\lambda_t\|^2}{\varepsilon^d} \right)^d
\]

\[
= d \ln \left( 1 + \frac{1}{\varepsilon} \sum_{t=1}^{T} \|\lambda_t\|^2_2 \right).
\]

### 6.2 Composite loss functions

In many (batch) statistical learning problems, we are interested in minimizing composite objectives such as

\[
\min_{w \in \mathbb{R}^d} \frac{1}{T} \sum_{t=1}^{T} \max\{0, 1 - y_t \langle x_t, w \rangle\} + \frac{\gamma}{2} \|w\|^2_2,
\]  

(SVM)

\[
\min_{w \in \mathbb{R}^d} \frac{1}{T} \sum_{t=1}^{T} (\langle x_t, w \rangle - y_t)^2 + \gamma \|w\|_1.
\]  

(Lasso)

Above, \(0.5\gamma \|w\|^2_2\) and \(\gamma \|w\|_1\) (with \(\gamma \geq 0\)) are convex penalty functions that correspond to a kind of problem-specific estimation bias. In the context of SVM, the \(0.5\gamma \|w\|^2_2\) penalty encourages half-space functions with large margins; in the context of Lasso, the \(\gamma \|w\|_1\) penalty encourages sparse weight vectors.

The general form of the batch convex optimization problem is

\[
\min_{w \in S} \sum_{t=1}^{T} (f_t(w) + g(w))
\]

where the \(f_t: \mathbb{R}^d \to \mathbb{R}\) are convex loss functions, and \(g: \mathbb{R}^d \to \mathbb{R}\) is a convex penalty function independent of \(t\). Loss functions of the form \(f_t + g\) are called composite loss functions.

In online convex optimization, we can handle convex composite loss functions just as before with previous algorithms, substituting \(f_t + g\) in place of \(f_t\) everywhere, and assuming that the feedback obtained in round \(t\) is
a subgradient $\lambda_t$ of the composite loss $f_t + g$ at $w_t$. However, this may lead to an undesirable algorithm. For instance, in the case of Lasso, where $g(w) = \gamma \|w\|_1$, the solution to the batch problem is often sparse (i.e., the minimizer has very few non-zero entries). On the other hand, the weight vector used by “Online gradient descent” (Algorithm 4) in round $t+1$, 

$$w_{t+1} = w_t - \eta \lambda_t,$$

is unlikely to be sparse unless both $w_t$ and $\lambda_t$ are sparse.

In order to take advantage of the composite nature of the loss functions, we will require a feedback model (in the protocol from Algorithm 1) slightly stronger than what we have been using so far. We assume that the non-time-varying part of the composite losses—i.e., the function $g$—is known (in some sense) to the learner, and further assume that in round $t$, the learner receives a subgradient $\lambda_t$ of $f_t$ at $w_t$. For instance, suppose it is easy to determine a subgradient of $g$ at any point in its domain. Then the learner can easily construct a subgradient of $f_t + g$ at $w_t$ under this feedback model, and hence execute algorithms such as Algorithm 6 and Algorithm 7. But it may be possible to do better using the known structure of $g$. In particular, when $\phi(w) := 0.5 \|w\|_2^2$, we may use the update rule

$$w_{t+1} := \arg \min_{w \in S} \eta \lambda_t, w) + \eta g(w) + \frac{1}{2} \|w - w_t\|_2^2. \quad (18)$$

This is almost the same as the update of Algorithm 4 (or Algorithm 7 with $\phi(w) = 0.5 \|w\|_2^2$), except that there, $g(w)$ is replaced with $\langle g_t, w \rangle$ for some $g_t \in \partial g(w_t)$ (i.e., in Algorithm 4 and Algorithm 7, both $f_t$ and $g$ are linearized, rather than just $f_t$). It turns out this update can be computed quite easily in many cases, and moreover it inherits desirable properties afforded by $g$ itself.

**Example 5.** For the Lasso problem, we use $g(w) := \gamma \|w\|_1$, $\phi(w) := 0.5 \|w\|_2^2$, and $S := \mathbb{R}^d$. In this case, the solution $w_{t+1}$ may be sparse on account of the sparsity-inducing property of the $l_1$ norm penalty. First, observe that (18) may be rewritten as

$$w_{t+1} := \arg \min_{w \in \mathbb{R}^d} \eta \gamma w + \frac{1}{2} \|\alpha - (w - \eta \lambda_t)\|_2^2,$$

(by completing the square). Let $\alpha := \eta \gamma$ (which we assume is positive) and $v := w_t - \eta \lambda_t$, so this is

$$w_{t+1} := \arg \min_{w \in \mathbb{R}^d} \sum_{i=1}^d \alpha |w_i| + \frac{1}{2} (w_i - v_i)^2.$$
We can therefore solve for each coordinate of $\mathbf{w}_{t+1}$ separately, so let us focus on a particular coordinate $i \in [d]$ and minimizing $w_i \mapsto \alpha|w_i| + 0.5(w_i - v_i)^2$. Per Proposition 3, $w_i \in \mathbb{R}$ is the minimizer if there exists a subgradient $\delta_i \in \partial|w_i|$ satisfying $\alpha \delta_i + w_i - v_i = 0$, i.e.,

$$w_i = v_i - \alpha \delta_i.$$ 

Recall that

$$\partial|w_i| = \begin{cases} 
\{-1\} & \text{if } w_i < 0, \\
[-1, +1] & \text{if } w_i = 0, \\
\{+1\} & \text{if } w_i > 0.
\end{cases}$$

Therefore the solution is as follows:

$$w_i = \begin{cases} 
v_i + \alpha < 0 & \text{if } v_i < -\alpha, \text{ with } \delta_i = -1, \\
0 & \text{if } -\alpha \leq v_i \leq \alpha, \text{ with } \delta_i = v_i/\alpha \in [-1, 1], \\
v_i - \alpha > 0 & \text{if } v_i > \alpha, \text{ with } \delta_i = 1.
\end{cases}$$

In other words, if $|v_i| \leq \alpha$, then it is set to zero ("killed"); if $|v_i| > \alpha$, it is shifted towards zero by an amount $\alpha$ ("shrunk").

**Example 6.** For the SVM problem, we use $g(\mathbf{w}) := 0.5\gamma\|\mathbf{w}\|_2^2$, $\phi(\mathbf{w}) := 0.5\|\mathbf{w}\|_2^2$, and $S := \mathbb{R}^d$. As per above, the update in (18) is

$$\mathbf{w}_{t+1} := \arg \min_{\mathbf{w} \in \mathbb{R}^d} \eta \gamma \|\mathbf{w}\|_2^2 + \frac{1}{2}\|\mathbf{w} - (\mathbf{w}_t - \eta \lambda_t)\|_2^2.$$

The solution is

$$\mathbf{w}_{t+1} = \frac{1}{1 + \eta \gamma} (\mathbf{w}_t - \eta \lambda_t).$$

### 6.3 Generalized online mirror descent

A general version of online mirror descent that handles both composite losses as well as time-dependent regularization functions is given in Algorithm 9.
**Algorithm 9** Generalized online mirror descent

**Input** Legendre regularization functions $\phi_t: \mathbb{R}^d \to \mathbb{R}$, convex penalty function $g: \mathbb{R}^d \to \mathbb{R}$, closed convex set $S \subseteq \text{int} \cap_{t \in [T]} \text{dom} \phi_t$.

1. Let $w_1 := \arg \min_{w \in S} \phi_1(w)$.
2. for $t = 1$ to $T$ do
   3. Receive feedback: $f_t(w_t)$ and $\lambda_t \in \partial f_t(w_t)$ for some convex function $f_t: \mathbb{R}^d \to \mathbb{R}$.
   4. Incur loss: $f_t(w_t) + g(w_t)$.
   5. Update:
      $$w_{t+1} := \arg \min_{w \in S} \langle \lambda_t, w \rangle + g(w) + D_{\phi_{t+1}}(w, w_t).$$  \hspace{1cm} (19)
3. end for

**Lemma 5.** Let Legendre functions $\phi_1, \phi_2, \ldots, \phi_T: \mathbb{R}^d \to \mathbb{R}$ be such that $\phi_t$ is strongly convex with respect to the norm $\| \cdot \|_{(t)}$ whose dual is $\| \cdot \|_{(t)+}$, and let $S \subseteq \text{int} \cap_{t \in [T]} \text{dom} \phi_t$ be a closed convex set. Let $f_1, f_2, \ldots, f_T: \mathbb{R}^d \to \mathbb{R}$ be any sequence of convex loss functions, and $g: \mathbb{R}^d \to \mathbb{R}$ be a convex penalty function. The learner using Algorithm 9 ("Generalized online mirror descent") satisfies the following. There exist subgradients $g'_t \in \partial g(w_t)$ such that for all $u \in S$,

$$\sum_{t=1}^T \langle \lambda_t, w_t - u \rangle + \langle g'_{t+1}, w_{t+1} - u \rangle \leq D_{\phi_1}(u, w_1) + \frac{1}{2} \sum_{t=1}^T \| \lambda_t \|_{(t+1)+}^2 + \sum_{t=1}^T (D_{\phi_{t+1}}(u, w_t) - D_{\phi_t}(u, w_t)).$$

**Proof.** By the optimality condition from Proposition 4 and the definition of $w_{t+1}$ in (19), there exists a subgradient $g'_{t+1} \in \partial g(w_{t+1})$ such that

$$\langle \lambda_t + g'_{t+1} + \nabla \phi_{t+1}(w_{t+1}) - \nabla \phi_{t+1}(w_t), w_{t+1} - u \rangle \leq 0$$  \hspace{1cm} (20)

for all $u \in S$. So fix a particular $u \in S$, and consider the following decomposition of $\langle \lambda_t, w_t - u \rangle + \langle g'_{t+1}, w_{t+1} - u \rangle$:

$$\langle \lambda_t, w_t - u \rangle + \langle g'_{t+1}, w_{t+1} - u \rangle = \langle \lambda_t, w_t - w_{t+1} \rangle + \langle \lambda_t + g'_{t+1}, w_{t+1} - u \rangle.$$
We bound the first term as follows:
\[
\langle \lambda_t, w_t - w_{t+1} \rangle \leq \|\lambda_t\|_{(t+1)^*} \|w_t - w_{t+1}\|_{(t+1)} \\
\leq \frac{1}{2} \|\lambda_t\|_{(t+1)^*}^2 + \frac{1}{2} \|w_t - w_{t+1}\|_{(t+1)}^2 \\
\leq \frac{1}{2} \|\lambda_t\|_{(t+1)^*}^2 + D_{\phi_{t+1}}(w_{t+1}, w_t). \tag{21}
\]

The first inequality follows by the definition of dual norm, the second inequality is by AM/GM, and the third inequality follows by strong convexity of $\phi_{t+1}$ with respect to the norm $\|\cdot\|_{(t+1)^*}$. To bound the second term, we use (20) to obtain
\[
\langle \lambda_t + g'_{t+1}, w_{t+1} - u \rangle \\
\leq \langle \nabla \phi_{t+1}(w_t) - \nabla \phi_{t+1}(w_{t+1}), w_{t+1} - u \rangle \\
= D_{\phi_{t+1}}(u, w_t) - D_{\phi_{t+1}}(u, w_{t+1}) - D_{\phi_{t+1}}(w_{t+1}, w_t) \\
= (D_{\phi_{t}}(u, w_t) - D_{\phi_{t+1}}(u, w_{t+1})) + (D_{\phi_{t+1}}(u, w_t) - D_{\phi_{t}}(u, w_t)) \\
- D_{\phi_{t+1}}(w_{t+1}, w_t). \tag{22}
\]
Combining (21) and (22), we have
\[
\langle \lambda_t, w_t - u \rangle + \langle g'_{t+1}, w_{t+1} - u \rangle \\
\leq \frac{1}{2} \|\lambda_t\|_{(t+1)^*}^2 + (D_{\phi_{t}}(u, w_t) - D_{\phi_{t+1}}(u, w_{t+1})) \\
+ (D_{\phi_{t+1}}(u, w_t) - D_{\phi_{t}}(u, w_t)).
\]
Summing this up over $t$ and dropping the non-positive term $-D_{\phi_{T+1}}(u, w_{T+1})$ gives the claimed bound. \hfill \Box

7 Boosting

In a typical binary classification problem, we are given training data $D := ((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)) \in (\mathcal{X} \times [-1, +1])^n$, and we want to find a relatively simple function (or “hypothesis”) $h: \mathcal{X} \rightarrow [-1, +1]$ such that its training error $\hat{err}(h) := \sum_{i=1}^n I\{h(x_i) \neq y_i\}/n$, is as small as possible.

A “weak learning oracle” takes as input the training data $D$ along with weights $w = (w_1, w_2, \ldots, w_n) \in \mathbb{R}^n$. It promises to return a hypothesis $h: \mathcal{X} \rightarrow [-1, +1]$ such that the 
edge or correlation of the hypothesis (with respect to $w$)
\[
\hat{\text{cor}}(h, w) := \sum_{i=1}^n w_i y_i h(x_i)
\]
is bounded away from zero. A boosting algorithm, over a sequence of rounds, adaptively chooses weights \( w_t \) and calls the weak learning oracle with \( w_t \) to get a hypothesis \( h_t \). After \( T \) rounds, the boosting algorithm returns a weighted combination of the hypotheses \( f := \sum_{t=1}^{T} \alpha_t h_t \), which is a function \( f : \mathcal{X} \to \mathbb{R} \); we typically threshold \( f \) by using \( h := \text{sign} \circ f \) (so \( h(x) = \text{sign}(f(x)) \)).

**Algorithm 10**  Boosting algorithm

**input**  Training data \( \mathcal{D} \), Legendre regularization function \( \phi : \mathbb{R}^n \to \mathbb{R} \), closed convex set \( S \subseteq \text{int dom } \phi \), initial weights \( w_1 \in S \), regularization parameter \( \eta > 0 \),

1:  for \( t = 1 \) to \( T \) do
2:     Call weak learning oracle with training data \( \mathcal{D} \) and weights \( w_t \) to obtain \( h_t : \mathcal{X} \to [-1, +1] \).
3:     Let \( \rho_t := (y_1 h_t(x_1), y_2 h_t(x_2), \ldots, y_n h_t(x_n)) \in [-1, 1]^n \).
4:     Let \( \gamma_t := \langle \rho_t, w_t \rangle \) and \( \lambda_t := \eta \gamma_t \rho_t \).
5:     Update:
     \[
     w_{t+1} := \arg \min_{w \in S} \eta \langle \lambda_t, w \rangle + D_\phi(w, w_t). \tag{23}
     \]
6:  end for
7:  return  \( f := \sum_{t=1}^{T} \gamma_t h_t \).

**Theorem 11.** Assume \( \phi \) is strongly convex with respect to \( \| \cdot \| \). Suppose Algorithm 10 is run with regularization parameter \( \eta > 0 \) such that \( \eta \| \rho_t \|_2^2 \leq 1 \) for all \( t = 1, 2, \ldots, T \). Then, for any \( u \in S \),

\[
\text{cor}(f, u) \geq \frac{1}{2} \sum_{t=1}^{T} \gamma_t^2 - \frac{D_\phi(u, w_1)}{\eta}.
\]

**Proof.** Algorithm 10 is simply Online mirror descent where the loss function in round \( t \) is \( \langle \lambda_t, \cdot \rangle \). From Theorem 9 for any \( u \in S \),

\[
\sum_{t=1}^{T} \langle \lambda_t, w_t - u \rangle \leq \frac{\eta}{2} \sum_{t=1}^{T} \| \lambda_t \|_2^2 + \frac{D_\phi(u, w_1)}{\eta}.
\]

Observe that

\[
\sum_{t=1}^{T} \langle \lambda_t, w_t \rangle = \sum_{t=1}^{T} \gamma_t^2 \quad \text{and} \quad \frac{\eta}{2} \sum_{t=1}^{T} \| \lambda_t \|_2^2 = \frac{\eta}{2} \sum_{t=1}^{T} \gamma_t^2 \| \rho_t \|_2^2 \leq \frac{1}{2} \sum_{t=1}^{T} \gamma_t^2.
\]
Moreover,
\[ \sum_{t=1}^{T} \langle \lambda_t, u \rangle = \sum_{i=1}^{n} u_i y_i \sum_{t=1}^{T} \gamma_t h_t(x_i) = \sum_{i=1}^{n} u_i y_i f_i(x_i) = \tilde{\text{cor}}(f, u). \]
Therefore
\[ \sum_{i=1}^{n} u_i y_i f(x_i) + \frac{D_\phi(u, w_1)}{\eta} \geq \frac{1}{2} \sum_{t=1}^{T} \gamma_t^2. \]

**Corollary 3.** Assume Algorithm 10 is run with \( \phi(w) := \sum_{i=1}^{n} w_i \ln w_i - \sum_{i=1}^{n} w_i \) (generalized negative entropy), \( S := \Delta^{n-1}, w_1 := 1/n, \) and \( \eta := 1. \) Let \( h := \text{sign} \circ f. \) Then
\[ \hat{\text{err}}(h) \leq \exp \left( -\frac{1}{2} \sum_{t=1}^{T} \gamma_t^2 \right). \]

**Proof.** Define the vector \( u = (u_1, u_2, \ldots, u_n) \in S \) by
\[ u_i := \frac{1}{n \cdot \hat{\text{err}}(h)} \cdot \mathbb{1}\{h(x_i) \neq y_i\} \quad \text{for each } i \in [n]. \]
Then
\[ \tilde{\text{cor}}(f, u) = \sum_{i=1}^{n} u_i y_i f(x_i) \leq 0 \quad \text{and} \quad D_\phi(u, w_1) = \ln \frac{1}{\text{err}(h)}. \]
The conclusion now follows from Theorem 11 after some rearranging. \( \square \)

**Bibliographic notes**

A gentle introduction to convex optimization is the freely available text by Boyd and Vandenberghe (2004). Two excellent treatments of convex optimization algorithms in the context of machine learning are the monographs of Shalev-Shwartz (2012) and Bubeck (2014). These notes are heavily based on these monographs. The original mirror descent algorithm is from Nemirovski and Yudin (1983), and the original dual averaging method is due to Nesterov (2009). Online Newton step is due to [Agarwal et al. (2007). The generalization of online mirror descent for composite losses is from Duchi et al. (2010). Boosting algorithms were first given by Schapire (1990) and Freund (1995), and the first practical boosting algorithm (“Adaboost”) was given by Freund and Schapire (1997); also see the text of Schapire and Freund (2012). Algorithm 10 is similar to Adaboost, but has a slightly different update step. The derivation of Algorithm 10 is adapted from Arora et al. (2012).
References


