The Hedge Algorithm on a Continuum Supplementary material, ICML 2015

Walid Krichene   Maximilian Balandat   Claire Tomlin
Alexandre Bayen

We present, for completeness, proofs which were omitted from the paper.

1 Proof of Lemma 4

Lemma 4. *If $\psi : L^2(S) \to \mathbb{R}$ is ℓ_ψ-strongly convex w.r.t. $\|\cdot\|$, then ψ^* is $\frac{1}{\ell_\psi}$ smooth w.r.t. $\|\cdot\|_*$, that is, for all x, y,*

$$\psi^*(x) - \psi^*(y) - \langle \nabla\psi^*(y), x - y\rangle \leq \frac{1}{2\ell_\psi}\|x - y\|_*^2$$

Proof. First, we prove that $\nabla\psi^*$ is $\frac{1}{\ell_\psi}$-Lipschitz (see for example Nesterov [2009]).

Let $y_1, y_2 \in E^*$, and $x_i = \nabla\psi^*(y_i)$. Since x_i is the minimizer of the convex function $x \mapsto \psi(x) - \langle y_i, x\rangle$, we have, by first-order optimality,

$$\langle \nabla\psi(x_i) - y_i, x - x_i\rangle \geq 0 \ \forall x \in \mathcal{X}$$

In particular, we have

$$\langle \nabla\psi(x_1) - y_1, x_2 - x_1\rangle \geq 0$$
$$\langle \nabla\psi(x_2) - y_2, x_1 - x_2\rangle \geq 0$$

and summing both inequalities,

$$\langle y_2 - y_1, x_2 - x_1\rangle \geq \langle \nabla\psi(x_2) - \nabla\psi(x_1), x_2 - x_1\rangle$$

By strong convexity, we have

$$\langle y_2 - y_1, x_2 - x_1\rangle = \langle \nabla\psi(x_2) - \nabla\psi(x_1), x_2 - x_1\rangle \geq \ell_\psi\|x_2 - x_1\|^2$$

and by definition of the dual norm, we have $\langle y_2 - y_1, x_2 - x_1\rangle \leq \|y_2 - y_1\|_*\|x_2 - x_1\|$. Therefore,

$$\|y_2 - y_1\|_*\|x_2 - x_1\| \geq \ell_\psi\|x_2 - x_1\|^2$$

rearranging, we have $\|x_2 - x_1\| \leq \frac{1}{\ell_\psi}\|y_2 - y_1\|_*$, i.e.

$$\|\nabla\psi^*(y_2) - \nabla\psi^*(y_1)\| \leq \frac{1}{\ell_\psi}\|y_2 - y_1\|_* \tag{1}$$

Finally,

$$\begin{aligned}
&\psi^*(x) - \psi^*(y) - \langle \nabla\psi^*(y), x - y\rangle \\
&\quad= \int_0^1 \langle \nabla\psi^*(y + t(x - y)) - \nabla\psi^*(y), x - y\rangle\, dt \\
&\quad\leq \|y - x\|_* \int_0^1 \|\nabla\psi^*(y + t(x - y)) - \nabla\psi^*(y)\| dt \\
&\quad\leq \|y - x\|_* \int_0^1 \frac{1}{\ell_\psi}\|y + t(x - y) - y\|_* dt && \text{by (1)} \\
&\quad\leq \frac{1}{\ell_\psi}\|x - y\|_*^2 \int_0^1 t dt \\
&\quad= \frac{1}{\ell_\psi}\|x - y\|_*^2 \frac{1}{2}
\end{aligned}$$

$\square$

2 Equivalence of Regret with respect to elements of S and elements of $\mathcal{X}$

In what follows, let $\mathcal{X} = \{f \in L^2(S) : f \geq 0 \text{ a.e. and } \int_S f(s)ds = 1\}$. Observe that $\mathcal{X}$ is closed: We have $\mathcal{X} = \mathcal{X}_1 \cap \mathcal{X}_2$, where $\mathcal{X}_1 = \{f \in L^2(S) : f \geq 0 \text{ a.e.}\}$ and $\mathcal{X}_2 = \{f \in L^2(S) : \int_S f(s)ds = 1\}$. $\mathcal{X}_1$ is clearly closed, and so is $\mathcal{X}_2$, being the inverse image of the closed set $\{1\}$ under the continuous mapping $f \mapsto \int_S f(s)ds$.

We show the equivalence between the regret with respect to elements of the set S and regret with respect to the set of Lebesgue continuous distributions on S, as stated formally in the following:

Suppose that the $\ell^{(\tau)}$ are L*-Lipschitz, uniformly in time, and that S is v-uniformly fat with respect to the Lebesgue uniform measure. Then*

$$\begin{aligned}
R^{(t)} &= \sum_{\tau=1}^{t} \left\langle \ell^{(\tau)}, x^{(\tau)} \right\rangle - \min_{s \in S} \sum_{\tau=1}^{t} \ell^{(\tau)}(s) \\
&= \sum_{\tau=1}^{t} \left\langle \ell^{(\tau)}, x^{(\tau)} \right\rangle - \inf_{x \in \mathcal{X}} \left\langle \sum_{\tau=1}^{t} \ell^{(\tau)}(s), x \right\rangle
\end{aligned}$$

Proof. Let $s_t^\star$ be a minimizer of $\sum_{\tau=1}^{t} \ell^{(\tau)}(s)$. Then it suffices to show that for all $\epsilon > 0$, there exists $x \in \mathcal{X}$ such that

$$\left\langle \sum_{\tau=1}^{t} \ell^{(\tau)}, x \right\rangle \leq \sum_{\tau=1}^{t} \ell^{(\tau)}(s_t^\star) + \epsilon$$

Fix $\epsilon > 0$. Since S is v-uniformly fat, there exists a convex set $K_t \subset S$ containing $s_t^\star$, with $\lambda(K_t) \geq v$. Let S_t be the homothetic transform of K_t as given in Lemma 3, of center $s_t^\star$ and ratio d_t yet to be determined. Then we have

$$\begin{aligned}
&D(S_t) = d_t D(K_t) \leq d_t D(S) \\
&\lambda(S_t) = d_t^n \lambda(K_t) > 0
\end{aligned}$$

Now consider $x = \frac{1}{\lambda(S_t)} 1_{S_t}$. We have $x \in \mathcal{X}$, and since the $\ell^{(\tau)}$ are uniformly L-Lipschitz,

$$\begin{aligned}
\left\langle \sum_{\tau=1}^{t} \ell^{(\tau)}, x \right\rangle &= \sum_{\tau=1}^{t} \int_{S_t} \frac{1}{\lambda(S_t)} \ell^{(\tau)}(s) ds \\
&\leq \sum_{\tau=1}^{t} \int_{S_t} \frac{1}{\lambda(S_t)} (\ell^{(\tau)}(s_t^\star) + \mathrm{L} d_t D(S)) ds \\
&= t\, \mathrm{L} d_t D(S) + \sum_{\tau=1}^{t} \ell^{(\tau)}(s_t^\star)
\end{aligned}$$

In particular, if we choose $d_t = \frac{\epsilon}{t \mathrm{L} D(S)}$, we have $\left\langle \sum_{\tau=1}^{t} \ell^{(\tau)}, x \right\rangle \leq \sum_{\tau=1}^{t} \ell^{(\tau)}(s_t^\star) + \epsilon$, which proves the claim. $\square$

3 Proof of Proposition 1

Next, we consider the dual averaging method when the regularization functional ψ is taken to be the negative entropy

$$\psi(x) = \int_S x(s) \ln x(s) ds + \lambda(S)$$

We prove Proposition 1, which show that the solution to the dual averaging iteration is given by the Hedge update rule:

Proposition 1. *Let $L^{(t)} \in E^*$, and consider the dual averaging iteration*

$$x^{(t+1)} \in \arg\min_{x \in \mathcal{X}} \left\langle L^{(t)}, x \right\rangle + \frac{1}{\eta_{t+1}} \psi(x) \tag{2}$$

where ψ is the negative entropy. Then the solution $x^{(t+1)}$ is given by the Hedge update rule:

$$x^{(t+1)}(s) = \frac{1}{\bar{Z}^{(t)}} e^{-\eta_{t+1} L^{(t)}(s)}$$

where $\bar{Z}^{(t)}$ is the normalization constant $\bar{Z}^{(t)} = \int_S e^{-\eta_{t+1} L^{(t)}(s)} ds$.

Proof. Let K be the cone $K = \{x \in L^2(S) : x \geq 0\}$, and let

$$f(x) = \left\langle L^{(t)}, x \right\rangle + \frac{1}{\eta_{t+1}} \psi(x) + i_K(x)$$

where i_K is the indicator function of the cone K, i.e. $i_K(s) = +\infty$ if $s \in K$ and 0 otherwise. The dual averaging iteration is equivalent to the following problem:

$$\begin{array}{ll} \text{minimize}_{x \in L^2(S)} & f(x) \\ \text{subject to} & \langle \mathbf{1}, x \rangle = 1 \end{array}$$

where $\mathbf{1} : S \to \mathbb{R}$ is identically equal to 1. Using the fact that the subdifferential of the indicator i_K is the normal cone N_K given by[1]

$$\forall x \in K,\ \partial i_K(x) = N_K(x) = \left\{ g \in L^2(S) : \sup_{y \in K} \langle g, y - x \rangle \leq 0 \right\},$$

the subdifferential of the objective function is

$$\partial f(x) = L^{(t)} + \frac{1}{\eta_{t+1}} (1 + \ln x) + N_K(x)$$

First, we show that, for all x and all $g \in N_K(x)$, $gx = 0$ almost everywhere. Indeed, fixing $x \in K$, we have $\langle g, y - x \rangle \leq 0$ for all $y \in K$. In particular, if we consider $y = x \left(1 + \frac{1}{2} 1_{g>0} - \frac{1}{2} 1_{g<0}\right)$, we have

$$\langle g, y - x \rangle = \left\langle g, x \left(\frac{1}{2} 1_{g>0} - \frac{1}{2} 1_{g<0} \right) \right\rangle = \frac{1}{2} \langle |g|, x \rangle = \frac{1}{2} \int_S |g(s)| x(s) ds$$

therefore $\frac{1}{2} \int_S |g(s)| x(s) ds \leq 0$, which implies that $|g|x = 0$ a.e..

Now, consider the Lagrangian $\mathcal{L} : E \times \mathbb{R} \to \mathbb{R}$

$$\mathcal{L}(x, \nu) = \left\langle L^{(t)}, x \right\rangle + \frac{1}{\eta_{t+1}} \psi(x) + i_K(x) + \nu(\langle \mathbf{1}, x \rangle - 1)$$

Then $(x^\star, \nu^\star)$ is an optimal pair only if

$$\begin{aligned} &0 \in L^{(t)} + \frac{1}{\eta_{t+1}} (1 + \ln x^\star) + N_K(x^\star) + \nu \mathbf{1} \\ &\langle \mathbf{1}, x^\star \rangle = 1 \end{aligned}$$

[1] See for example Chapter 16 in Bauschke and Combettes [2011]

see for example Bauschke and Combettes [2011] Section 19.3. We can rewrite the stationarity condition in the following way:

$$\exists\, g^\star \in N_K(x^\star) \text{ such that } L^{(t)} + \frac{1}{\eta_{t+1}}(1 + \ln x^\star) + \nu \mathbf{1} + g^\star = 0.$$

Therefore,

$$\begin{aligned}
&x^\star(s) = e^{-\eta_{t+1} L^{(t)}(s)} / e^{1+\eta_{t+1}(\nu^\star + g^\star(s))} \text{ a.e.} \\
&g^\star \in N_K(x^\star) \\
&\langle \mathbf{1}, x^\star \rangle = 1
\end{aligned}$$

In particular, $x^\star > 0$ a.e., thus by the observation that $g^\star x^\star = 0$ a.e., we must have $g^\star = 0$ a.e. Therefore, the necessary conditions become

$$\begin{aligned}
&x^\star(s) = \frac{e^{-\eta_{t+1} L^{(t)}(s)}}{\bar{Z}^{(t)}} \\
&\bar{Z}^{(t)} = e^{1+\eta_{t+1}\nu^\star} \\
&\frac{\int e^{-\eta_{t+1} L^{(t)}(s)} ds}{\bar{Z}^{(t)}} = 1
\end{aligned}$$

which proves the claim. □

References

H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer, 2011.

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.