Some calculations about the mirror descent (MD) algorithm.

The setting and notations...

- The potential $R : \mathbb{R}^d \to \mathbb{R}$, a Legendre function.
- Its convex conjugate: $R^*(u) = \sup_x \{\langle u, x\rangle - R(x)\}$.
- The "mirror maps" $\nabla R : \mathbb{R}^d \to \mathbb{R}^d$ and $\nabla R^* : \mathbb{R}^d \to \mathbb{R}^d$.
- The loss functions $(\ell_t)$, each $\ell_t : \mathbb{R}^d \to \mathbb{R}$.
- Time index $t \in \{0, 1, 2, \dots\}$ for discrete time, $t \in [0, \infty)$ for continuous time.
- $x$, $x_t$ (discrete time), $x(t)$ (continuous time) elements of the primal space.
- $u$, $u_t$ (discrete time), $u(t)$ (continuous time) elements of the dual space.

The MD algorithm (discrete time) uses the initialization and updates

$$
\begin{aligned}
x_1 &= \arg\min R(x)\,, \\
x_{t+1} &= \arg\min_{x\in\mathbb{R}^d} \Big\{ \langle \nabla \ell_t(x_t), x\rangle + \eta D_R(x, x_t) \Big\}\,.
\end{aligned}
$$

Here $D_R(x, y)$ stands for the Bregman divergence, whose definition is recalled now:

$$D_R(a_1, a_2) = R(a_1) - R(a_2) - \langle \nabla R(a_2), a_1 - a_2\rangle\,.$$

The order of the arguments matters. Let's say $D_R(a_1, a_2)$ is "supported at $a_2$" (its definition uses a tangent supported at $a_2$). To minimize clutter let's denote by $\partial_i$ the partial derivative (operator) with respect to the $i$-th argument. So for instance:

$$
\begin{aligned}
\partial_1 D_R(a_1, a_2) &= \nabla R(a_1) - \nabla R(a_2)\,, \\
\partial_2 D_R(a_1, a_2) &= -\nabla^2 R(a_2)(a_1 - a_2)\,.
\end{aligned}
$$

A couple of equations in discrete time:

$$u_{t+1} = u_t - \eta \nabla \ell_t(x_t) \tag{1dt}$$

$$x_t = \nabla R^*(u_t)\,. \tag{2dt}$$

Analogous equations in continuous time:

$$u'(t) = -\eta \nabla \ell_t(x(t)) \tag{1ct}$$

$$x(t) = \nabla R^*(u(t))\,. \tag{2ct}$$

For fixed $x^* \in \mathbb{R}^d$, $u^* = \nabla R(x^*)$,

$$\ell_t(x^*) \geq \ell_t(x(t)) + \langle \nabla \ell_t(x(t)), x^* - x(t)\rangle \tag{3}$$

From (1ct) we get $\nabla \ell_t(x(t)) = -u'(t)/\eta$, and plugging this in (3) we get:

$$\ell_t(x^*) \geq \ell_t(x(t)) - \frac{1}{\eta}\langle u'(t), x^* - x(t)\rangle \,.$$

Rearranging:

$$\ell_t(x(t)) - \ell_t(x^*) \leq \frac{1}{\eta}\langle u'(t), x^* - x(t)\rangle \,, \tag{4}$$

and the claim is that the RHS is the derivative of a Bregman divergence.

On one hand, by the chain rule and the symmetry of the Hessian:

$$\begin{aligned}
\frac{d}{dt} D_{R^*}(u^*, u(t)) &= \langle \partial_2 D_{R^*}(u^*, u(t)), u'(t)\rangle \\
&= -\langle \nabla^2 R^*(u(t))(u^* - u(t)), u'(t)\rangle \\
&= -\langle u^* - u(t), \nabla^2 R^*(u(t)) u'(t)\rangle \\
&= -\langle u^* - u(t), x'(t)\rangle \,.
\end{aligned}$$

On the other hand,

$$\frac{d}{dt}\langle u^* - u(t), x^* - x(t)\rangle = -\langle u'(t), x^* - x(t)\rangle - \langle u^* - u(t), x'(t)\rangle \,.$$

Combining the two:

$$\begin{aligned}
\langle u'(t), x^* - x(t)\rangle &= -\langle u^* - u(t), x'(t)\rangle - \frac{d}{dt}\langle u^* - u(t), x^* - x(t)\rangle \\
&= \frac{d}{dt} D_{R^*}(u^*, u(t)) - \frac{d}{dt}\langle u^* - u(t), x^* - x(t)\rangle \,.
\end{aligned}$$

Then (4) can be written as

$$\ell_t(x(t)) - \ell_t(x^*) \leq \frac{1}{\eta}\frac{d}{dt}\Big(D_{R^*}(u^*, u(t)) - \langle u^* - u(t), x^* - x(t)\rangle\Big)$$

After integrating (???)

$$\int_0^T [\ell_t(x(t)) - \ell_t(x^*)] dt \leq \frac{1}{\eta}\big[D_{R^*}(u^*, u(T)) - D_{R^*}(u^*, u(0))\big]$$

Note: the term $\langle u^* - u(t), x^* - x(t)\rangle = \langle \nabla R(x^*) - \nabla R(x(t)), x^* - x(t)\rangle$ is the sum of $D_R(x^*, x(t))$ and $D_R(x(t), x^*)$, so it is non-negative.