Bregman Divergences

On Bregman Divergences

Definition:

The Bregman Divergence is defined as:

$$B_R(x, y) = R(x) - R(y) - \nabla R(y)^T (x - y)$$

Clearly, $B_R(x, y) = 0$ if $x = y$. For R a convex function, $B_R(x, y) \geqslant 0$. For R a strictly convex function, $B_R(x, y) > 0$ unless $x = y$.

Main Property:

OCO uses the following property of Bregman divergences quite heavily. A "strongly convex" function R is one for which $B_R(x, y) \geqslant \frac{\|x-y\|^2}{2}$ (for some norm $\|\cdot\|$). We have the following inequality for a strongly convex function, and this is (pretty much) the only place where the mystifying *Hardy's flop* is used. Note the following identity first:

$$B_R(x, y) + B_R(y, x) = [\nabla R(x) - \nabla R(y)] \circ (x - y)$$

This for instance is a special case of the 4-point equality, or can be derived directly. Now, note that this is expressing $B_R(x, y)$ as an inner product with $(x - y)$ (neglecting the other B_R term). So if we assume that R is strongly convex, we have that $B_R(x, y) \geqslant \frac{\|x-y\|^2}{2}$, and a similar bound for $B_R(y, x)$. We use the fact that $B_R(x, y) \geqslant 0$ for all x, y. Let the dual norm of $\|\cdot\|$ be $\|\cdot\|_\star$. Thus we have that:

$$\begin{aligned} [\nabla R(x) - \nabla R(y)] \circ (x - y) &\geqslant B_R(x, y) \\ &\geqslant \frac{\|x - y\|^2}{2} \end{aligned}$$

which, thanks to Holder's inequality yields:

$$\|\nabla R(x) - \nabla R(y)\|_\star \cdot \|x - y\| \geqslant \frac{\|x - y\|^2}{2}$$

i.e.

$$2 \|\nabla R(x) - \nabla R(y)\|_\star \geqslant \|x - y\|$$

or that,

$$2 \|\nabla R(x) - \nabla R(y)\|_\star^2 \geqslant B_R(x, y) \text{ for any } x, y$$

Of course if we were to include the $B_R(y, x)$ term too, then we would get a slightly tighter bound (getting rid of the factor of 2 in the above). Thus, we have the following lemma:

Lemma 1. *Suppose we are given a regularizer R that is strongly convex with regard to the norm $\|\cdot\|$, with dual norm $\|\cdot\|_\star$. Then it holds that:*

$$\|\nabla R(x) - \nabla R(y)\|_\star^2 \geqslant B_R(x, y) \text{ for all } x, y$$

Underlying Intuition:

What is a strongly convex function R? The technical description is one whose Bregman divergence is lower bounded by the square of some norm. The intuitive feel is this: the graph of the function is suitably curved, i.e. strong convexity captures a notion of curvature. Another way of describing this is that the ***double derivative*** of such a function R is *bounded away* from 0. Now look at the inequality

$$2\,\|\nabla R(x) - \nabla R(y)\|_\star \;\geqslant\; \|x - y\|$$

in this new perspective. What it means is that some quantity that "looks" like a double derivative

$$\frac{\|\nabla R(x) - \nabla R(y)\|_\star}{\|x - y\|} \;\geqslant\; \frac{1}{2}$$

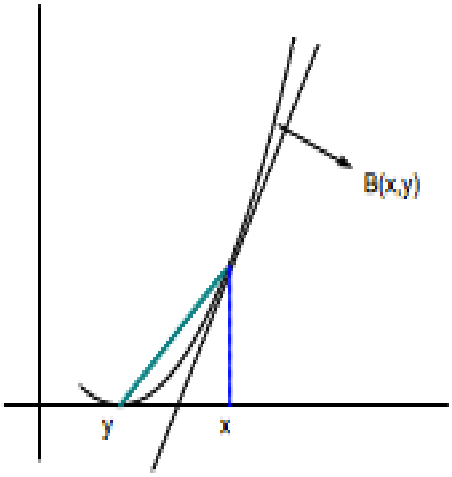
so this is indeed expected. The norm and the dual norm appear naturally in the above calculations when we try to make this notion of the double derivative (or in case of multivariate functions, the Hessian of the function R) rigorous.

So in a way, the usage of *Hardy's flop* is just restating that the function R is actually "curved".

We are also going to re-interpret the computation that says

$$[\nabla R(x) - \nabla R(y)] \circ (x - y) \;\geqslant\; B_R(x, y)$$

Think of this in the following manner. Let $B_R(x, y)$ be viewed as a function of x (i.e. fix a value of y). Thus, $B_R(x, y)$ is a convex function and takes on the value 0 at the point y. Note that the gradient of this function at the point x is precisely $[\nabla R(x) - \nabla R(y)]$. Given this, the inequality stated above is essentially restating the fact that $B_R(x,\ y)$ is a *convex* function in the (first) parameter x. The inequality above says that the gradient of $B_R(x, y)$ (the **black** solid tangent in the figure below) has a ***higher*** slope than the secant formed (the **green** line in the figure below) between the points x and y on the curve $B_R(x, y)$.



A Generalization of the Main Property:

We have the following:

Lemma 2. ***(Projection Lemma)*** *Let y_1 and y_2 be two points and K a convex body. Let R be a regularizer function that is strongly convex with respect to the norm $\|\cdot\|$, and let x_1 and x_2 be the Bregman projections of y_1 and y_2 respectively, on to the convex body K. Then the following statement holds:*

$$\|x_1 - x_2\| \leqslant \|\nabla R(y_1) - \nabla R(y_2)\|_\star$$

*where $\|\cdot\|_\star$ is the **dual** norm of $\|\cdot\|$.*

Why is this a generalization of the main property? Just consider $y_1 = x_1$ and $y_2 = x_2$. In this case, the projection is a no-op.

Proof. The proof is quite similar to that of the earlier result. For sake of tedium, let us repeat it. Since x_1 is the Bregman projection of y_1 on the convex body K, we have that the gradient of the Bregman divergence at the point x_1 cannot have any direction of decrease that leads inside the convex body K, in particular in the direction $x_2 - x_1$. Thus,

$$[\nabla R(x_1) - \nabla R(y_1)] \circ (x_2 - x_1) \;\geqslant\; 0$$

Also, similarly, considering the points x_2 and y_2 and letting x_1 be the "other" point in K, we have:

$$[\nabla R(x_2) - \nabla R(y_2)] \circ (x_1 - x_2) \;\geqslant\; 0$$

Adding the two inequalities and a little manipulation gives us:

$$\begin{aligned}[\nabla R(y_1) - \nabla R(y_2)] \circ (x_1 - x_2) &\geqslant [\nabla R(x_1) - \nabla R(x_2)] \circ (x_1 - x_2) \\ &\geqslant \|x_1 - x_2\|^2\end{aligned}$$

and a final application of Holder's inequality gives us the result. This is in fact Hardy's flop in action again.

Also note that this gives us that

$\|\nabla R(y_1) - \nabla R(y_2)\|_\star^2 \geqslant [\nabla R(y_1) - \nabla R(y_2)] \circ (x_1 - x_2) \geqslant B_R(x_1, x_2)$

in case we wanted a form similar to the statement in Lemma 1.

□

Comments:

We have asked later whether it is possible to prove $\|y_1 - y_2\| \geqslant \|x_1 - x_2\|$. Now it seems that such a thing may not be possible to prove. Of course counterexamples may be hard to find, but the basic philosophy is that "duality" is quite strong an indicator of what can or cannot hold. Note the form of the inequality above. While one side is in the primal norm, the other side is in the dual norm. This does not seem to be artefact of the analysis but rather a form that is indeed required.

Also of course, deriving intuition from the 2-norm case is not very wise, since for the 2-norm, we have that the dual norm is the primal norm itself, and also because $\nabla R(y) = y$. So the above also re-proves the "Projection Lemma" in Hazan Kale, for the specific case of the 2-norm.

While we are not able to prove $\|y_1 - y_2\| \geqslant \|x_1 - x_2\|$, in most cases, we can safely assume that this is the case and that will help us understand the calculations a lot better (for instance, it takes away the need to go through the pesky application of Hardy's flop). Of course, at the end, we will have to re-do the correct calculations using the flop as above.

The Active Update:

The active update is given by

$$\begin{aligned} \nabla R(y_{t+1}) &= \nabla R(x_t) - \eta c_t \\ x_{t+1} &= \text{argmin}_{x \in K} B_R(x, y_{t+1}) \end{aligned}$$

Based on the above, we can quickly analyse the update as follows. The t^{th} regret term is as follows:

$$\begin{aligned} c_t \circ (x_t - u) &= [\nabla R(x_t) - \nabla R(y_{t+1})] \circ (x_t - u)/\eta \\ &= [B_R(u, x_t) - B_R(u, y_{t+1}) + B_R(x_t, y_{t+1})]/\eta \\ &= [B_R(u, x_t) - B_R(u, y_{t+1})]/\eta + B_R(x_t, y_{t+1})/\eta \\ &\leqslant [B_R(u, x_t) - B_R(u, x_{t+1})]/\eta + B_R(x_t, y_{t+1})/\eta \end{aligned}$$

In the second line, we use the 4-point equality. The last step is because of Pythagoras, that x_{t+1} is the projection of y_{t+1}. From the foregoing, we can upperbound $B_R(x_t, y_{t+1})$:

$$\begin{aligned} B_R(x_t, y_{t+1}) &\leqslant \|\nabla R(x_t) - \nabla R(y_{t+1})\|_\star^2 \\ &= \|\eta c_t\|_\star^2 \end{aligned}$$

so,

$$\begin{aligned} c_t \circ (x_t - u) &\leqslant [B_R(u, x_t) - B_R(u, x_{t+1})]/\eta + \|\eta c_t\|_\star^2/\eta \\ &= [B_R(u, x_t) - B_R(u, x_{t+1})]/\eta + \eta \|c_t\|_\star^2 \end{aligned}$$

Now the rest of the regret calculation is clear: we have to sum this up over $t = 1...T$ and set $\eta = \frac{1}{\sqrt{T}}$, and we are done.

The Lazy Update:

Interestingly the lazy update is slightly trickier than the active update. The update states:

$$\begin{aligned} \nabla R(y_{t+1}) &= \nabla R(y_t) - \eta c_t \\ x_{t+1} &= \text{argmin}_{x \in K} B_R(x, y_{t+1}) \end{aligned}$$

As before, the t^{th} regret term is as follows:

$$\begin{aligned} \eta c_t \circ (x_t - u) &= [\nabla R(y_t) - \nabla R(y_{t+1})] \circ (x_t - u) \\ &= B_R(u, y_t) - B_R(u, y_{t+1}) + B_R(x_t, y_{t+1}) - B_R(x_t, y_t) \\ &= [B_R(u,\, y_t) - B_R(x_t,\, y_t)] - [B_R(u,\, y_{t+1}) - B_R(x_{t+1},\, y_{t+1})] + [B_R(x_t,\, y_{t+1}) - \\ &\quad B_R(x_{t+1}, y_{t+1})] \end{aligned}$$

The second equality above follows from the 4-point equality; in the third step, we add and subtract a term $B_R(x_{t+1}, y_{t+1})$. Why do we do this?

Although the term $B_R(u,\, y_t) - B_R(u,\, y_{t+1})$ looks like it can be telescoped, that would not serve our purpose, since the end result of the telescoping would give us $B_R(u,\, y_1)$. We do not have much control over this in terms of the *diameter* of the convex body K. This is why, we are having to consider the expression $B_R(u,\, y_t) - B_R(x_t,\, y_t)$. Even this expression does not look like it can be bounded by the diameter, but interestingly, at the boundary, the telescoping sum gives us $B_R(u, y_1) - B_R(x_1, y_1)$. Here, note that y_1 is the unconstrained minimizer of the R function, which makes $\nabla R(y_1) = 0$. Thus, the expression

$$B_R(u, y_1) - B_R(u, x_1) = R(u) - R(x_1)$$

Here, both u and x_1 belong to the convex body K, and this can be thought of as upper bounded by the *diameter* of the convex body K (as measured by the strongly convex function R). This is the *importance* of choosing the *first* point y_1 as the *minimizer* of the function R.

And now the fun starts. We will set $u{=}x_{t+1}$ in the above to derive:

$$\begin{aligned}\eta c_t \circ (x_t - x_{t+1}) &= [B_R(x_{t+1}, y_t) - B_R(x_t, y_t)] + [B_R(x_t, y_{t+1}) - B_R(x_{t+1}, y_{t+1})] \\ &\geqslant B_R(x_{t+1}, x_t) + B_R(x_t, x_{t+1}) \\ &\geqslant \|x_t - x_{t+1}\|^2\end{aligned}$$

Here, we used Pythagoras in the second inequality step and noted that x_t is the projection of y_t and x_{t+1} is the projection of y_{t+1}. Again, let's think through this. Why do we set $u{=}x_{t+1}$? Well, if we were to set $u{=}x_t$, then we would get the useless tautology, $0{=}0$. It will not be good behaviour to set $u{=}y_t$ or $u{=}y_{t+1}$ because u is meant to be inside K and y_t, y_{t+1} are not necessarily inside K.

Continuing our argument: by the flop, this indicates that

$$\|\eta c_t\|_\star^2 \geqslant [B_R(x_t, y_{t+1}) - B_R(x_{t+1}, y_{t+1})]$$

and now we are done. We have a telescoping sum for the first part in the expression for $\eta c_t \circ (x_t - u)$, and the second part is upperbounded by $\eta^2 \|c_t\|_\star^2$.

Note that here, we could not apply a modularized version of the flop, like we could use for the active update. The calculations are all the more beautiful because there were so many ways to go wrong. In the case of the active update, the calculations were relatively straightforward.

Questions:

About projections:

Suppose y_1 and y_2 are two points in space, and K be a convex body. Let R be a strongly convex regularizer (will just convex do?) and let x_1 and x_2 be the projections of y_1 and y_2 respectively on K. Then can we claim the following:

$$B_R(x_1, x_2) \leqslant B_R(y_1, y_2)?$$

Failing this, given that R is strongly convex wrt a norm $\|\cdot\|$, is it true that

$$\|x_1 - x_2\| \leqslant \|y_1 - y_2\|?$$

Is it also true for instance that $\|x_1 - x_2\| \leqslant \|y_1 - x_2\|$ etc.?

There is a version of this as **Projection Lemma** (Lemma 9) in Hazan Kale's JMLR COLT08 paper on gradual variations. I do believe that this is true, we have to search the literature or work up a proof. Given that, even for the lazy update, we can do away with Hardy's flop as:

$$\eta c_t \circ (x_t - x_{t+1}) \geqslant [B_R(x_t, y_{t+1}) - B_R(x_{t+1}, y_{t+1})]$$

and by the foregoing,

$$\begin{aligned}\|\eta c_t\|_\star \cdot \|y_t - y_{t+1}\| &\geqslant \|\eta c_t\|_\star \cdot \|x_t - x_{t+1}\| \\ &\geqslant [B_R(x_t, y_{t+1}) - B_R(x_{t+1}, y_{t+1})]\end{aligned}$$

and we are done with the bound.

About a partial order:

Suppose we are given two points x and y. Under what conditions does it hold that $B_R(x,y) > B_R(y,x)$. If we define a relation based on this $x \succ y$ iff $B_R(x,y) > B_R(y,x)$, then is this relation an order?

About the definition of a strong convex function:

The definition of a strongly convex function R involves a norm $\|\cdot\|$ and we have that $B_R(x,y) \geqslant \frac{\|x-y\|^2}{2}$. The question is this: what happens if we define a strongly convex function R with regard to a norm $\|\cdot\|$ as one that satisfies $B_R(x,y) \geqslant \frac{\|x-y\| \cdot \|x-y\|_\star}{2}$. Note that for the usual 2-norm the definition stays the same as the old definition.

Proof of Pythagoras Theorem for Bregman Projections:

Given a convex body K and a point y outside the convex body (in the ambient space), the Bregman projection of y, with respect to the regularizer R, is defined as

$$x \;=\; \text{argmin}_{u \in K} B_R(u,y)$$

Observe that this is a constrained optimization problem. If we had an unconstrained optimization problem minimizing a function ϕ at a point x, then we have that $\nabla\phi(x) = \vec{0}$, i.e. the gradient at the point x vanishes. The proof of this essentially consists of the following argument: if ϕ had a non-vanishing gradient at x, then one could step in (the opposite of) that direction and lower ϕ even more, by the definition of what a gradient is. Thus, the gradient of ϕ cannot have any non-vanishing component at the point x. For a constrained problem, the core argument still stays the same: ϕ is minimum at a point x of K implies that it cannot have any negative component in any direction that goes inside the convex body K. For if so, then we could "walk" in the direction of that component and achieve a smaller value of the function (while staying inside the convex body). For our problem, $\nabla B_R(u,y) = \nabla R(u) - \nabla R(y) = \nabla R(x) - \nabla R(y)$, at the point x, and so we have:

$$[\nabla R(x) - \nabla R(y)] \circ (u - x) \;\geqslant\; 0$$

Apply the 4-point equality to this statement and we have:

$$B_R(x,x) + B_R(u,y) - B_R(u,x) - B_R(x,y) \;\geqslant\; 0$$

i.e.

$$B_R(u,y) \;\geqslant\; B_R(u,x) + B_R(x,y)$$

and this is precisely Pythagoras' theorem for Bregman projections.

A modification of Pythagoras':

The following lemma gets used in the gradual variations paper by Mahdavi et al. (COLT 2012). Let

$$v \;=\; \text{argmin}_{x \in K} [\ell \circ x + B_R(x,u)]$$

where ℓ is a specific cost vector (so that $\ell \circ x$ is a *linear* cost function). What can we say about v from the gradient information of the objective function? The lemma is the statement marked **red** below. The idea is the same as before. If the objective function is minimized at $v \in K$, then it has to be true that the gradient of the objective function at the minimizer point v does not have any directions that "go inside" the convex body and give an improvement to the objective function. In this case the gradient of the objective function at the point v is $[\ell + \nabla R(v) - \nabla R(u)]$. Given any other point $w \in K$, there should not be any improvement in the direction of $w - v$. Thus, we will have:

$$[\ell + \nabla R(v) - \nabla R(u)] \circ (w - v) \;\geqslant\; 0$$

Or in other words,

$$\begin{aligned} [\nabla R(v) - \nabla R(u)] \circ (w - v) &\geqslant \ell \circ (v - w) \\ i.e., B_R(v, v) + B_R(w, u) - B_R(w, v) - B_R(v, u) &\geqslant \ell \circ (v - w) \\ i.e., B_R(w, u) - B_R(w, v) - B_R(v, u) &\geqslant \ell \circ (v - w) \end{aligned}$$

Let's record this as a

Lemma 3. *Suppose we are given a strongly convex regularizer R, and a vector ℓ. Also suppose v is the following minimizer:*

$$v = \text{argmin}_{x \in K}[\ell \circ x + B_R(x, u)]$$

Then it holds that (for any $w \in K$):

$$B_R(w, u) - B_R(w, v) - B_R(v, u) \geqslant \ell \circ (v - w)$$

An application of the above:

Suppose we consider the usual (active) update in the following form:

$$x_{t+1} = \text{argmin}_{x \in K}[c_t \circ x + B_R(x, x_t)]$$

Then applying the above lemma gives us that for any $w \in K$:

$$B_R(w, x_t) - B_R(w, x_{t+1}) - B_R(x_{t+1}, x_t) \geqslant c_t \circ (x_{t+1} - w)$$

The usual interpretation for this is that "Prescient Marginal Cost" is always low. In this case, the choice of x_{t+1} is made after the cost function c_t comes along, and so we expect that *prescient* regret to be low. Indeed it is, as can be seen from the above that nicely forms a telescoping sum.

An idle calculation:

Here, we want to check out the ramifications of the following fact. If y is fixed, then we can consider the function $B_R(x, y)$ just as a function of x – this can be thought of as the "distance like measure" from the point y. Note that the derivative/gradient of this function at the point x is nothing but $\nabla B_R(x, y) = \nabla R(x) - \nabla R(y)$. Now, go back to all of the above calculations and restate all of the above in terms of the gradient of the Bregman divergence to see if some more intuition results.

Other Properties:

The Bregman Divergence $B_R(x, y)$ is not symmetric.

Also, $B_R(x, y)$ is *convex* in the *first* argument x but not in the argument y. This is easy to see: fix a y in the above definition, and then you see that $B_R(x, y)$ is then a convex function $R(x)$ along with an affine function $\nabla R(y)^T x$ (if we ignore all the constant terms that depend only on the fixed y). Thus, overall, B_R is convex in x.

This manifests itself as follows: when we do a Bregman projection, we perform $\text{argmin}_{x \in K} B_R(x, y)$. The minimization is done with regard to the first argument, because then this is a meaningful minimization (minimizing a convex function).

Duality:

The Bregman divergence function enjoys the following duality relationship. Given a function R, consider the Fenchel dual of R, and call it $R^\star$. Then we have that

$$B_R(x, y) = B_{R^\star}(y^\star, x^\star)$$

where $y^\star$ and $x^\star$ are the "duals" of y and x respectively (with respect to R) i.e. $y^\star = \nabla R(y)$ and similarly for $x^\star$. This proof is simply a matter of writing out the definitions of B_R and the Fenchel dual of a function R.

The 4-point equality for Bregman Divergences:

In the following, we will derive the four point equality of Bregman Divergences. In a manner that is not entirely formal, we can draw connections between Bregman Divergences and the "energy" of a system - it is essentially a second order quantity. This will come out in various inequalities involving Bregman Divergences, like for instance the Pythagorean Identity for Bregman Divergences.

Now, consider the expression $[\nabla R(a) - \nabla R(b)] \circ (c - d)$. For the time being we will forget the first order terms (i.e. $R(a)$, $R(b)$ etc.). Then note that, $\nabla R(y) \circ (y - x)$ can serve as a *proxy* for $B_R(x, y)$. Now we start moulding the above expression:

$$\nabla R(a) \circ (c - d) = \nabla R(a) \circ (a - d) - \nabla R(a) \circ (a - c) \approx B_R(d, a) - B_R(c, a)$$

And a similar expression for $\nabla R(b) \circ (c - d)$. Overall, we thereby get the four point equality for Bregman Divergences (and here notice that the first order terms cancel out):

$$[\nabla R(a) - \nabla R(b)] \circ (c - d) = B_R(d, a) + B_R(c, b) - B_R(c, a) - B_R(d, b)$$

The broad motto can be the following: Since Bregman Divergences are *second order* quantities, we can expect some identity involving two terms one involving ∇'s (i.e. $\nabla R(a) - \nabla R(b)$) and the other term involving just a linear expression (i.e. $(c - d)$). How do we remember this equality?

Consider the following identity:

$$(a - b)(c - d) = ac + bd - bc - ad$$

An alternate way of writing this by always placing the entities in the second bracket first is:

$$(a - b)(c - d) = ca + db - cb - da$$

Finally, the irritating thing to remember is that the minuses are the "+" terms in the 4-point equality and vice versa.

Of course, setting (say, $b = c$) in the 4-point equality, we get the 3-point equality:

$$[\nabla R(a) - \nabla R(b)] \circ (b - d) = B_R(d, a) - B_R(b, a) - B_R(d, b)$$

which may be rewritten as

$$[\nabla R(a) - \nabla R(b)] \circ (d - b) = B_R(d, b) + B_R(b, a) - B_R(d, a)$$

The duality view to Online Convex Optimization:

In the following, we will discuss the **Fenchel** dual paper by Shalev-Schwartz and Singer. We do not expect the dual to spit out the intuition that we need to keep a regularizer for online convex optimization. So the overall objective of the offline (one-shot) adversary (including the regularizer) is the following:

$\sum_{t=1}^{T} f_t(w) + \frac{R(w)}{\eta}$, where the adversary gets the single shot w (and where w is constrained to lie in the convex body K). This is the precise modified function that we minimize to get the prediction in the $t+1^{\text{st}}$ iteration, in FTRL. Now, we can "extend" this formulation out in order to bring in more constraints (which we will then form the Lagrangian dual of). We can have time-indexed variables w_t and with constraints $w_t = w_0$ for every t. Then the modified problem stands:

$$\begin{aligned} \text{min.} \quad & \sum_{t=1}^{T} f_t(w_t) + \frac{R(w_0)}{\eta} \\ s.t. \quad & w_0 \in K \\ \forall t \geqslant 1: \quad & w_t = w_0 \end{aligned}$$

Clearly, this models the original adversary. Now we pull the constraints $w_t = w_0$ into the objective function, keeping a Lagrangian multiplier λ_t. The Lagrangian multiplier λ_t is actually a vector (since it is associated with the vector equality $w_t = w_0$). The multipliers correspond to equality constraints, so do not have any restriction on them. Thus, the Lagrangian expression reads:

$$\begin{aligned} \text{min.} \quad & \sum_{t=1}^{T} f_t(w_t) + \frac{R(w_0)}{\eta} + \lambda_t \circ (w_0 - w_t) \\ s.t. \quad & w_0 \in K \end{aligned}$$

(We will later investigate as to why we kept $w_t = w_0$. Another option would have been to keep $w_t = w_{t-1}$ for all t.)

The objective function in this last convex program nicely separates out into terms corresponding to w_0 and w_t (the problems are separable). Thus, we have the following re-writing of the program above:

$$\begin{aligned} \text{min.} \quad & \sum_{t=1}^{T} [f_t(w_t) - \lambda_t \circ w_t] + \left[\left(\sum_{t=1}^{T} \lambda_t \right) \circ w_0 + \frac{R(w_0)}{\eta} \right] \\ s.t. \quad & w_0 \in K \end{aligned}$$

The Lagrangian dual will consist of the above expression (including the min.) maximized over all possible λ_t's (the "best" lower bound). But for any fixed λ, the above expression separates nicely, and we start seeing the Fenchel dual arising. The Fenchel dual is

$$f^\star(x) \;=\; \max.(w \circ x - f(w))$$

so,

$$\min.(f(w) - w \circ x) \;=\; -f^\star(x)$$

A final re-writing of the above program in terms of the Fenchel dual then leads to the following Lagrangian dual:

$$\max._{\lambda_t} \quad \left[\left\{ -\sum_{t=1}^{T} f_t^\star(\lambda_t) \right\} - \frac{R^\star\left(-\eta \sum_{t=1}^{T} \lambda_t\right)}{\eta} \right]$$

So far so good. We will now proceed to set up the online framework, and then we will get some sanity checks on what values of λ_t should we set.

In the online framework, we have cost functions f_t ($t = 1 \cdots T$) coming along, and we make the prediction w_t before the t^{th} cost function f_t comes along. When the t^{th} cost function comes along, we will do **two** things: we will set the λ_t in the dual convex program above, and we will set the new prediction w_{t+1} in the primal convex program.

First, imagine that the f_t's are actually linear functions. This is okay because of the usual linearization step involved in OCO. Given that, however, the $f_t^\star$ has a very specific form. To be precise, $f_t^\star(w)$ is $+\infty$ unless $w = c_t$ where c_t is *the* gradient of the cost function f_t. Thus, given this understanding, we would need to set $\lambda_t = c_t$ or else the above maximization problem would give us a value of $-\infty$ which is not at all good from the point of view of a suitable lower bound on the primal problem. Thus, in the t^{th} round, after we have already predicted w_t, we receive the cost function f_t, and we consider the gradient of f_t at the point w_t. This is what we should set λ_t as.

How do we predict w_{t+1}? The algorithm sets w_{t+1} according to the gradient of the dual objective function; noting that $f_t^\star(\lambda_t)$ is 0, we have that

$$w_{t+1} \;=\; \nabla R^\star\left(-\eta \sum_{j=1}^{t} \lambda_j\right)$$

(In case this is not familiar, we have the following factoid: if $\nabla R(w) = z$, then $w = \nabla R^\star(z)$. So this above is just a restatement of the following

$$\nabla R(w_{t+1}) \;=\; -\eta \sum_{j=1}^{t} \lambda_j$$

Now, given that we are going to set $\lambda_j = c_j$, this fits into what we already know about the FTRL update.)

What I do not understand is that the gradient is the direction of sharpest increase, but the term above appears with a negative sign in the dual objective function, so why do we go in the direction of the gradient? Does this have to do with the fact that R is *strongly convex*, so that $R^\star$ is *smooth*?

Before we proceed, let us quickly check the fact mentioned above.

Lemma 4. *If R is a convex function, and $\nabla R(w) = z$, then $w = \nabla R^\star(z)$*

Proof. Note that $R^\star(z) = \max_w \{w \circ z - R(w)\}$ in the unconstrained world, so that the maximum is achieved when the gradient of the expression inside the bracket vanishes. Thus, $z = \nabla R(w)$ for the w that achieves the maximum. Now we can use the fact that if $R^\star$ is the Fenchel conjugate of R, then R is the Fenchel conjugate of $R^\star$. This happens under certain general conditions, if R is convex and closed for instance. Thus, we automatically get that $w = \nabla R^\star(z)$.

□

On Gradual Variations

In the following, we will apply the facts above to analyse the gradual variations paper of Mahdavi et al [COLT 2012]. Let us recapitulate the broad theme of OCO in the following terms. The prediction x_t is made first, and then the cost function c_t comes along that dictates the penalty of the prediction. However, c_t is indeed taken into account when computing the prediction x_{t+1}. Thus, the marginal *prescient* regret $c_t \circ (x_{t+1} - u)$ (for any fixed u) can be expected to be **low**. And most calculations indicate that this is indeed the case.

Broadly a regret calculation follows the following steps. Let us informally make the following

Definition 5. *We call an expression a **good expression** if it can be readily seen as something that is easily bounded by the diameter of the convex set K (as measured by the strongly convex regularizer R) or the gradient of the cost functions (perhaps in the dual norm).*

For instance, the foregoing paragraph may be summarized as follows: the expression $c_t \circ (x_{t+1} - u)$ is a ***good expression*** (for any $u \in K$ and any t where $1 \leqslant t \leqslant T$).

Lemma 6. *The expression $c_t \circ (x_t - x_{t+1})$ is a **good expression** in typical OCO settings, when the active or the lazy update is considered.*

Proof. We will assume that R is a strongly convex regularizer. For instance, given the lazy update, we have that $c_t = [\nabla R(y_t) - \nabla R(y_{t+1})]/\eta$. Now, we can apply the **Projection Lemma** (Lemma 2), to get that:

$$\begin{aligned} \eta \|c_t\|_\star &= \|\nabla R(y_t) - \nabla R(y_{t+1})\|_\star \\ &\geqslant \|x_t - x_{t+1}\| \end{aligned}$$

Thus, we have that

$$\begin{aligned} c_t \circ (x_t - x_{t+1}) &\leqslant \|c_t\|_\star \|x_t - x_{t+1}\| \\ &\leqslant \eta \|c_t\|_\star^2 \end{aligned}$$

This implies that the expression $c_t \circ (x_t - x_{t+1})$ is a ***good expression***.

Consider the active update, where $c_t = [\nabla R(x_t) - \nabla R(y_{t+1})]/\eta$. Again, we can apply Lemma 2, and imagine the points x_t and $y_{t+!}$ being projected to x_t and $x_{t+!}$ to derive the inequality

$$\|\nabla R(x_t) - \nabla R(y_{t+1})\|_\star \geqslant \|x_t - x_{t+1}\|$$

and again we are done.

□

Given all these facts about ***good expressions***, we can have a 1-line derivation of "low regret":

$$c_t \circ (x_t - u) = c_t \circ (x_t - x_{t+1}) + c_t \circ (x_{t+1} - u)$$

Thus, $c_t \circ (x_t - u)$ is the sum of 2 ***good expressions***, and we are done.

In Mahdavi et al., they want to upper bound the regret in terms of the variations of the cost functions, i.e. some cumulative measure of the vectors $(c_t - c_{t-1})$. Suppose we were to perform the usual (say, active) update (for now, like in their paper, we are letting $\eta = 1$).

$$\begin{aligned} c_t &= \nabla R(x_t) - \nabla R(y_{t+1}) \\ c_{t-1} &= \nabla R(x_{t-1}) - \nabla R(y_t) \end{aligned}$$

We then compute $c_t - c_{t-1}$ and get an expression with 4 terms:

$$c_t - c_{t-1} = \nabla R(x_t) - \nabla R(x_{t-1}) - \nabla R(y_{t+1}) + \nabla R(y_t)$$

This is too complicated to handle, and may not yield what we want. **The main theme going through these papers is that, in order to keep the expressions wieldy, only allow two ∇s to enter any expression.** Ostensibly so that we could use the 4-point equality, etc. Supposing that this is the guiding principle, how do we proceed? The thing is that when we receive c_t we can use x_t to make the next update. But, we can also sort of "ignore" the most recent prediction x_t and instead use x_{t-1}. How does this help? Let us say that for every **even** t, we perform the update:

$$c_t = \nabla R(x_t) - \nabla R(y_{t+1})$$

and for every **odd** t, we perform:

$$c_t \;=\; \nabla R(x_{t-1}) - \nabla R(y_{t+1})$$

Then notice what happens when t is **odd**:

$$\begin{aligned} c_t &= \nabla R(x_{t-1}) - \nabla R(y_{t+1}) \\ c_{t-1} &= \nabla R(x_{t-1}) - \nabla R(y_t) \end{aligned}$$

Now, when we consider the expression $c_t - c_{t-1}$, the terms corresponding to x_{t-1} vanish, and we are left with:

$$c_t - c_{t-1} \;=\; \nabla R(y_t) - \nabla R(y_{t+1})$$

But now we have that (by the **Projection Lemma**, i.e. Lemma 2):

$$\|c_t - c_{t-1}\|_\star \;\geqslant\; \|x_t - x_{t+1}\|$$

for such an update rule (of course only when t is **odd**). Thus this shows that when t is **odd**, the expression $(c_t - c_{t-1}) \circ (x_t - x_{t+1})$ is a ***good expression***.

When t is **even**, a similar rule will work, but where, for **even** t, we perform the update:

$$c_t \;=\; \nabla R(x_{t-1}) - \nabla R(y_{t+1})$$

We first note the identity:

$$\begin{aligned} c_t \circ (x_t - u) &= c_t \circ (x_t - x_{t+1}) + c_t \circ (x_{t+1} - u) \\ &= c_t \circ (x_t - x_{t+1}) + \text{good expression} \\ &= (c_t - c_{t-1}) \circ (x_t - x_{t+1}) + c_{t-1} \circ (x_t - x_{t+1}) + \text{good expression} \\ &= (c_t - c_{t-1}) \circ (x_t - x_{t+1}) + \text{good expression} + \text{good expression} \\ &= \text{good expression} + \text{good expression} + \text{good expression} \end{aligned}$$

Here, the regret bound has terms for $\|c_t - c_{t-1}\|_\star^2$ and we have a regret bound for gradually varying cost functions.

Now how do we reconcile both of these updates? That is where simple "**juxtaposition**" works. What does this mean? This means that in each update, we have *two* descriptions for the cost function c_t. Playing around with the descriptions slightly, we land upon Mahdavi et al's updates.

It turns out that Mahdavi et al. uses the active form of the updates. One could also use a *lazy* form of update and a similar result holds. Typically, the active update is preferable since the calculations for the "prescient marginal cost" are comparatively simpler than for the lazy update.

Open Questions:

For Mahdavi et al, why can we not just make the update depend instead of c_t on $c_t - c_{t-1}$? Thus, what is the problem with the following update:

$$\nabla R(y_{t+1}) \;=\; \nabla R(y_t) - (c_t - c_{t-1})$$

If we could do this, then we would directly get a regret bound in terms of $(c_t - c_{t-1})$? Notice, however that if the cost functions do not vary, then we have that $y_{t+1} = y_t$, i.e. the prediction does not vary at all, which should be fine too.

Thereby it is quite clear why "two" Bregman projections are needed for getting a regret bound in terms of gradual variations. Can we consider the variational quantity $(c_t - 2c_{t-1} + c_{t-2})$ and show how we need to consider *three* Bregman projections etc.?

What insight does this give us for the original Buchbinder Naor problem?

Can we give a primal dual analysis for the case of gradual variations? At least for the case of unconstrained optimization? This would be the Shalev-Schwartz & Singer parallel for the regret bound for gradual variations. For this, look at the Fenchel dual that we discussed last time (when we discussed the Shalev-Schwartz & Singer paper) – in that Fenchel dual, we took the equality constraints as $x_t = x_0$ and we added these in the Lagrangian dual. One direction we did not explore is if we keep constraints $x_t = x_{t-1}$ instead, in the Lagrangian dual, we get terms for $(c_t - c_{t-1})$. Can this be worked into a regret bound in terms of gradual variations?

Failing progress on the Buchbinder Naor paper for general convex bodies, general norms, we have another question: modify the regret bounds in Buchbinder Naor, so that it pertains to gradual variations. This would imply getting new update rules, perhaps involving "two" projections to the probability simplex. For this, a starting point is to look at Mahdavi's paper for the section on multiplicative weight updates.