ONLINE OPTIMIZATION HW-3

SIDDHANT CHAUDHARY
BMC201953

All the problems below are taken from Chapter 5 from Elad Hazan's book.

Problem 3 of Chapter 5. Let $R(x) = \frac{1}{2}||\boldsymbol{x} - \boldsymbol{x}_0||^2$. We will show that the Bregman divergence corresponding to R is the Euclidean metric.

First, observe that for any $\boldsymbol{y}$, we have

$$\nabla R(\boldsymbol{y}) = \boldsymbol{y} - \boldsymbol{x}_0$$

So, by the definition of Bregman divergence, we have the following.

$$\begin{aligned}
B_R(\boldsymbol{x}||\boldsymbol{y}) &= R(\boldsymbol{x}) - R(\boldsymbol{y}) - \nabla R(\boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y}) \\
&= \frac{1}{2}||\boldsymbol{x} - \boldsymbol{x}_0||^2 - \frac{1}{2}||\boldsymbol{y} - \boldsymbol{x}_0||^2 - (\boldsymbol{y} - \boldsymbol{x}_0)^T(\boldsymbol{x} - \boldsymbol{y}) \\
&= \frac{1}{2}||\boldsymbol{x} - \boldsymbol{x}_0||^2 - \frac{1}{2}||\boldsymbol{y} - \boldsymbol{x}_0||^2 - (\boldsymbol{y} - \boldsymbol{x}_0)^T(\boldsymbol{x} - \boldsymbol{x}_0 - (\boldsymbol{y} - \boldsymbol{x}_0)) \\
&= \frac{1}{2}||\boldsymbol{x} - \boldsymbol{x}_0||^2 + \frac{1}{2}||\boldsymbol{y} - \boldsymbol{x}_0||^2 - (\boldsymbol{y} - \boldsymbol{x}_0)^T(\boldsymbol{x} - \boldsymbol{x}_0) \\
&= \frac{1}{2}||\boldsymbol{x} - \boldsymbol{x}_0 - (\boldsymbol{y} - \boldsymbol{x}_0)||^2 \\
&= \frac{1}{2}||\boldsymbol{x} - \boldsymbol{y}||^2
\end{aligned}$$

And this is what we wanted to show.

Now, recall that the projection with respect to this divergence is defined to be the quantity

$$\operatorname*{argmin}_{\boldsymbol{x} \in \mathcal{K}} B_R(\boldsymbol{x}||\boldsymbol{y})$$

Hence, in our case, the projection with respect to the divergence is

$$\operatorname*{argmin}_{\boldsymbol{x} \in \mathcal{K}} \frac{1}{2}||\boldsymbol{x} - \boldsymbol{y}||^2 = \operatorname*{argmin}_{\boldsymbol{x} \in \mathcal{K}} ||\boldsymbol{x} - \boldsymbol{y}||^2 = \operatorname*{argmin}_{\boldsymbol{x} \in \mathcal{K}} ||\boldsymbol{x} - \boldsymbol{y}||$$

and this is nothing but the standard Euclidean projection. This completes the solution to the problem.

Problem 5 of Chapter 5. For this problem, let us set up some notation. Let $\mathbf{1}$ denote the all ones vector, i.e all the coordinates of this vector are 1. For any vector $\boldsymbol{z}$, let $\log \boldsymbol{z}$ denote the vector in which we have applied the logarithm function to each coordinate of $\boldsymbol{z}$.

Let $\mathcal{K}$ be the n-dimensional simplex. Let $R(\boldsymbol{x}) = \boldsymbol{x}^T \log \boldsymbol{x}$ be the negative entropy regularization function. Computing the gradient of R, we get the following.

$$\nabla R(\boldsymbol{y}) = \mathbf{1} + \log \boldsymbol{x}$$

Date: 11 December, 2021.

Then the Bregman divergence is the following.

$$\begin{aligned} B_R(\boldsymbol{x}||\boldsymbol{y}) &= R(\boldsymbol{x}) - R(\boldsymbol{y}) - \nabla R(\boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y}) \\ &= \boldsymbol{x}^T \log \boldsymbol{x} - \boldsymbol{y}^T \log \boldsymbol{y} - (\mathbf{1} + \log \boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y}) \\ &= \boldsymbol{x}^T \log \boldsymbol{x} - \boldsymbol{y}^T \log \boldsymbol{y} - \mathbf{1}^T(\boldsymbol{x} - \boldsymbol{y}) - \boldsymbol{x}^T \log \boldsymbol{y} + \boldsymbol{y}^T \log \boldsymbol{y} \\ &= \boldsymbol{x}^T(\log \boldsymbol{x} - \log \boldsymbol{y}) - \mathbf{1}^T(\boldsymbol{x} - \boldsymbol{y}) \end{aligned}$$

So, we conclude that the Bregman divergence is simply the relative entropy plus an additional term. But in our case, note that because $\boldsymbol{x}, \boldsymbol{y}$ are in the n-simplex, we have that $\mathbf{1}^T\boldsymbol{x} = \mathbf{1}^T\boldsymbol{y} = 1$. So, it follows that

$$B_R(\boldsymbol{x}||\boldsymbol{y}) = \boldsymbol{x}^T(\log \boldsymbol{x} - \log \boldsymbol{y})$$

and hence $B_R(\boldsymbol{x}||\boldsymbol{y})$ is indeed the relative entropy.

Now, we will show that D_R, the diameter of $\mathcal{K}$ with respect to R, satisfies the upper bound $D_R^2 \leq \log n$. The proof is pretty simple. First, note that by definition, we have

$$D_R^2 = \max_{\boldsymbol{x},\boldsymbol{y} \in \mathcal{K}} R(\boldsymbol{x}) - R(\boldsymbol{y}) = \max_{\boldsymbol{x},\boldsymbol{y} \in \mathcal{K}} \sum_{i=1}^n x_i \log x_i - \sum_{i=1}^n y_i \log y_i$$

Now, we focus on the quantity

$$\sum_{i=1}^n x_i \log x_i - \sum_{i=1}^n y_i \log y_i$$

Because $0 \leq x_i \leq 1$, we see that the sum $\sum_{i=1}^n x_i \log x_i \leq 0$. Infact, this sum is zero if $\boldsymbol{x}$ is a vertex of $\mathcal{K}$. So, it follows that maximizing the above quantity is the same as maximizing the quantity

$$-\sum_{i=1}^n y_i \log y_i = \sum_{i=1}^n y_i \log \frac{1}{y_i}$$

over $\mathcal{K}$. Now, note that the function $f(x) = \log x$ is *concave*. So, by *Jensen's Inequality* for concave functions, we know that if $\boldsymbol{y} = (y_1, ..., y_n) \in \mathcal{K}$, then

$$f\left(y_1 \cdot \frac{1}{y_1} + \cdots + y_n \cdot \frac{1}{y_n}\right) \geq y_1 f\left(\frac{1}{y_1}\right) + \cdots + y_n f\left(\frac{1}{y_n}\right)$$

The above inequality implies that

$$\log n \geq y_1 \log \frac{1}{y_1} + \cdots + y_n \log \frac{1}{y_n}$$

Ofcourse, above we assumed that all y_is are non-zero. Even if some of them are zeros, applying the same trick gives us an even stronger upper bound. So, putting everything above together, we see that

$$\sum_{i=1}^n x_i \log x_i - \sum_{i=1}^n y_i \log y_i \leq \log n$$

Infact, the above bound is tight; take $\boldsymbol{x}$ to be a vertex of $\mathcal{K}$, and let $\boldsymbol{y}$ be the uniform distribution. In that case, the first quantity is 0 and the second quantity is $\log n$. This shows that $D_R^2 \leq \log n$.

Finally, we show that projections with respect to this divergence over the simplex amounts to scaling by the ℓ_1 norm. So let $\boldsymbol{y}$ be any point with positive coordinates (we need this because we take the logarithm of $\boldsymbol{y}$ in the Bregman divergence). As we've calculated the Bregman divergence above, the projection of the point $\boldsymbol{y}$ onto the simplex $\mathcal{K}$ is the following.

$$\begin{aligned}\underset{\boldsymbol{x}\in\mathcal{K}}{\operatorname{argmin}} &= \boldsymbol{x}^T(\log\boldsymbol{x} - \log\boldsymbol{y}) - \mathbf{1}^T(\boldsymbol{x}-\boldsymbol{y}) \\ &= \sum_{i=1}^n x_i \log\frac{x_i}{y_i} - \sum_{i=1}^n x_i + \sum_{i=1}^n y_i \\ &= \sum_{i=1}^n x_i \log\frac{x_i}{y_i} - 1 + \sum_{i=1}^n y_i\end{aligned}$$

So, minimizing the above quantity is equivalent to minimizing the sum

$$\sum_{i=1}^n x_i \log\frac{x_i}{y_i}$$

Consider the function $f(x) = x\log x$, which we know is convex. Also note that

$$\begin{aligned}\sum_{i=1}^n x_i \log\frac{x_i}{y_i} &= \sum_{i=1}^n y_i \frac{x_i}{y_i}\log\frac{x_i}{y_i} \\ &= \sum_{i=1}^n y_i f\left(\frac{x_i}{y_i}\right) \\ &= ||\boldsymbol{y}||_1 \sum_{i=1}^n \frac{y_i}{||\boldsymbol{y}||_1} f\left(\frac{x_i}{y_i}\right)\end{aligned}$$

Now, by *Jensen's Inequality* for convex functions, we have the following.

$$\begin{aligned}||\boldsymbol{y}||_1 \sum_{i=1}^n \frac{y_i}{||\boldsymbol{y}||_1} f\left(\frac{x_i}{y_i}\right) &\geq ||\boldsymbol{y}||_1 f\left(\sum_{i=1}^n \frac{x_i}{||\boldsymbol{y}||_1}\right) \\ &= ||\boldsymbol{y}||_1 f\left(\frac{1}{||\boldsymbol{y}||_1}\right) \\ &= \log\frac{1}{||\boldsymbol{y}||_1}\end{aligned}$$

Moreover, it can be observed that $\boldsymbol{x} = \frac{\boldsymbol{y}}{||\boldsymbol{y}||_1}$ achieves the above minimum value. So, it follows that the projection with respect to this Bregman divergence of $\boldsymbol{y}$ onto the simplex is just $\frac{\boldsymbol{y}}{||\boldsymbol{y}||_1}$, which shows that these projections just amount to scaling by the ℓ_1 norm. This completes the solution of the problem.

Problem 10 of Chapter 5. First, let $A \succeq B \succ 0$ be two positive definite matrices. We show that $A^{\frac{1}{2}} \succeq B^{\frac{1}{2}}$. Before proving this, we prove a simple lemma.

Lemma 0.1. *Let $M \succeq 0$ be any positive semi-definite matrix. If N is any matrix, then $N^T MN \succeq 0$. The inequality is strict if in addition it is assumed that $M \succ 0$ and N is invertible.*

Proof. It is clear that $N^T MN$ is symmetric, because M is symmetric. Next, suppose $\boldsymbol{x}$ is some vector. Then, observe that

$$\boldsymbol{x}^T(N^T MN)\boldsymbol{x} = (N\boldsymbol{x})^T M(N\boldsymbol{x}) \geq 0$$

because M is positive semi-definite. Clearly, if N is invertible and $M \succ 0$, the inequality is actually strict. This completes the proof. ■

Now, coming back to the main problem, we know that $A - B \succeq 0$. By **Lemma** 0.1, and using the fact that $B^{-1/2}$ is a symmetric matrix (because $B^{1/2}$ is), we see that

$$B^{-1/2}AB^{-1/2} - I = B^{-1/2}(A - B)B^{-1/2} \succeq 0$$

By the same lemma (**Lemma** 0.1), $B^{-1/2}AB^{-1/2}$ is a positive definite matrix (since A is); infact, by the above inequality, we see that all eigenvalues of $B^{-1/2}AB^{-1/2}$ are greater than 1. Moreover, the above inequality implies that for all $\boldsymbol{x}$ such that $||\boldsymbol{x}|| = 1$,

$$\left\langle B^{-1/2}AB^{-1/2}\boldsymbol{x}, \boldsymbol{x}\right\rangle \geq \langle I\boldsymbol{x}, \boldsymbol{x}\rangle = 1$$

Next, we will use the simple identity

$$\langle A\boldsymbol{x}, \boldsymbol{y}\rangle = \left\langle \boldsymbol{x}, A^T\boldsymbol{y}\right\rangle$$

for any matrix A and vectors $\boldsymbol{x}, \boldsymbol{y}$. For any vector $\boldsymbol{x}$ such that $||\boldsymbol{x}|| = 1$, we have the following.

$$\left\langle B^{-1/2}AB^{-1/2}\boldsymbol{x}, \boldsymbol{x}\right\rangle \geq \langle I\boldsymbol{x}, \boldsymbol{x}\rangle = 1 \tag{0.1}$$

$$\implies \left\langle AB^{-1/2}\boldsymbol{x}, (B^{-1/2})^T\boldsymbol{x}\right\rangle \geq 1 \tag{0.2}$$

$$\implies \left\langle A^{1/2}B^{-1/2}\boldsymbol{x}, (A^{1/2})^T(B^{-1/2})^T\boldsymbol{x}\right\rangle \geq 1 \qquad (A = A^{1/2}A^{1/2}) \tag{0.3}$$

$$\implies \left\langle A^{1/2}B^{-1/2}\boldsymbol{x}, A^{1/2}B^{-1/2}\boldsymbol{x}\right\rangle \geq 1 \qquad (A^{1/2}, B^{1/2} \text{ are symmetric}) \tag{0.4}$$

$$\implies \left|\left|A^{1/2}B^{-1/2}\boldsymbol{x}\right|\right| \geq 1 \tag{0.5}$$

Now, consider the matrix $A^{1/2}B^{-1/2}$. Note that

$$B^{-1/4}A^{1/2}B^{-1/4} = B^{-1/4}(A^{1/2}B^{-1/2})B^{1/4}$$

and this implies that $A^{1/2}B^{-1/2}$ is similar to the matrix $B^{-1/4}A^{1/2}B^{-1/4}$; this means that they have the same eigenvalues. But, note that $B^{-1/4} = (B^{-1/4})^T$ (it is symmetric), and hence by **Lemma** 0.1, we have that $B^{-1/4}A^{1/2}B^{-1/4} \succ 0$ (because $A^{1/2} \succ 0$), and hence all eigenvalues of this matrix are positive (and real). Moreover, inequality (0.5) implies that all eigenvalues of $A^{1/2}B^{-1/2}$ are greater than 1 in absolute value; so it follows that all eigenvalues of $B^{-1/4}A^{1/2}B^{-1/4}$ are greater than one. This implies

$$B^{-1/4}A^{1/2}B^{-1/4} - I \succeq 0$$

By **Lemma** 0.1, we see that

$$B^{1/4}(B^{-1/4}A^{1/2}B^{-1/4} - I)B^{1/4} \succeq 0$$

and clearly this implies that $A^{1/2} - B^{1/2} \succeq 0$, and this proves our claim.