We would like to thank all reviewers for insightful comments. In the response below, we address the main concern of Reviewer 2. Due to the space limit, we cannot address other comments, which are indeed interesting and helpful. All typos and suggestions related to the presentation and the writing of the paper will be fixed/included in the final version.

Response to Reviewer 2. The main concern is whether there is a bug in the proof of the online stochastic mirror descent, specifically the inequality between line 254 and line 255 in the supplementary material.
In fact, there is **no** bug in the proof.

Let's consider first the given example (copy below).

```
- take Phi to be 0.5*ell_2^2
- take all norms to be ell_2
- take theta_t=y_{t}=(0,1)
- take theta_{t+1}=y_{t+1} = (0,0)
- set K to be half plane where first coordinate is bigger than say 1
- thus x_t =(1,1)
- the inequality does not hold, as you would need 1 >= 2
```

By the choice of $\Phi(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, we have $\nabla\Phi(\boldsymbol{x}^t) = \boldsymbol{x}^t$. By our notation, $\boldsymbol{\theta}^t = \nabla\Phi(\boldsymbol{x}^t)$, so $\boldsymbol{\theta}^t = \boldsymbol{x}^t$. Consequently, we do not understand why in the example, $\boldsymbol{\theta}^t$ is taken to be $(0,1)$ and $\boldsymbol{x}^t$ can be $(1,1)$.

Besides, between line 254 and line 255 (in the supplementary material, i.e., full paper), $\boldsymbol{\theta}^{t+1}$ does not involve so we do not understand the role of $\boldsymbol{\theta}^{t+1}$ in the example here. We try to guess whether the reviewer meant $\boldsymbol{\vartheta}^{t+1}$. However, even with that guess, we do not see any contradiction.

In the following, we give the proof of the inequality between line 254 and 255 with very detail explanation. Recall that $\Psi^t = \frac{1}{\eta}D_\Phi(\boldsymbol{x}^*\|\boldsymbol{x}^t)$. First, we observe that

$$\begin{aligned}
\eta\left(\Psi^{t+1} - \Psi^t\right) &= D_\Phi(\boldsymbol{x}^*\|\boldsymbol{x}^{t+1}) - D_\Phi(\boldsymbol{x}^*\|\boldsymbol{x}^t) && (1)\\
&\leq D_\Phi(\boldsymbol{x}^*\|\boldsymbol{y}^{t+1}) - D_\Phi(\boldsymbol{x}^*\|\boldsymbol{x}^t) && (2)\\
&= \Phi(\boldsymbol{x}^*) - \Phi(\boldsymbol{y}^{t+1}) - \langle \underbrace{\nabla\Phi(\boldsymbol{y}^{t+1})}_{\boldsymbol{\vartheta}^{t+1}}, \boldsymbol{x}^* - \boldsymbol{y}^{t+1}\rangle - \Phi(\boldsymbol{x}^*) + \Phi(\boldsymbol{x}^t) + \langle \underbrace{\nabla\Phi(\boldsymbol{x}^t)}_{\boldsymbol{\theta}^t}, \boldsymbol{x}^* - \boldsymbol{x}^t\rangle && (3)\\
&= \Phi(\boldsymbol{x}^t) - \Phi(\boldsymbol{y}^{t+1}) - \langle \boldsymbol{\vartheta}^{t+1}, \boldsymbol{x}^t - \boldsymbol{y}^{t+1}\rangle - \langle \boldsymbol{\vartheta}^{t+1} - \boldsymbol{\theta}^t, \boldsymbol{x}^* - \boldsymbol{x}^t\rangle && (4)\\
&= \Phi(\boldsymbol{x}^t) - \Phi(\boldsymbol{y}^{t+1}) - \langle \boldsymbol{\theta}^t, \boldsymbol{x}^t - \boldsymbol{y}^{t+1}\rangle + \langle \eta\boldsymbol{g}^t, \boldsymbol{x}^t - \boldsymbol{y}^{t+1}\rangle + \langle \eta\boldsymbol{g}^t, \boldsymbol{x}^* - \boldsymbol{x}^t\rangle && (5)\\
&\leq -\frac{\alpha_\Phi}{2}\|\boldsymbol{y}^{t+1} - \boldsymbol{x}^t\|^2 + \eta\langle \boldsymbol{g}^t, \boldsymbol{x}^t - \boldsymbol{y}^{t+1}\rangle + \eta\langle \boldsymbol{g}^t, \boldsymbol{x}^* - \boldsymbol{x}^t\rangle && (6)\\
&\leq \frac{\eta^2}{2\alpha_\Phi}\|\boldsymbol{g}^t\|_*^2 + \eta\langle \boldsymbol{g}^t, \boldsymbol{x}^* - \boldsymbol{x}^t\rangle && (7)
\end{aligned}$$

where

(1) by definition of Ψ^t;

(2) by the generalized Pythagorean property (Lemma 1);

(3) by the definition of the Bregman divergence;

(4) by notation $\boldsymbol{\vartheta}^{t+1} = \nabla\Phi(\boldsymbol{y}^{t+1})$ and $\boldsymbol{\theta}^t = \nabla\Phi(\boldsymbol{x}^t)$;

(5) using $\boldsymbol{\vartheta}^{t+1} = \boldsymbol{\theta}^t - \eta\cdot\boldsymbol{g}^t$ by the algorithm;

(6) using the α_Φ-strong convexity of Φ, specifically, $\Phi(\boldsymbol{x}^t)-\Phi(\boldsymbol{y}^{t+1})-\langle\boldsymbol{\theta}^t, \boldsymbol{x}^t-\boldsymbol{y}^{t+1}\rangle \leq -\frac{\alpha_\Phi}{2}\|\boldsymbol{y}^{t+1}-\boldsymbol{x}^t\|^2$ since $\Phi(\boldsymbol{y}^{t+1}) \geq \Phi(\boldsymbol{x}^t) + \langle\boldsymbol{\theta}^t, \boldsymbol{y}^{t+1} - \boldsymbol{x}^t\rangle + \frac{\alpha_\Phi}{2}\|\boldsymbol{y}^{t+1} - \boldsymbol{x}^t\|^2$ where recall $\boldsymbol{\theta}^t = \nabla\Phi(\boldsymbol{x}^t)$ and $-\langle\boldsymbol{\theta}^t, \boldsymbol{x}^t - \boldsymbol{y}^{t+1}\rangle = \langle\boldsymbol{\theta}^t, \boldsymbol{y}^{t+1} - \boldsymbol{x}^t\rangle$;

(7) using Cauchy-Schwarz inequality $\langle\boldsymbol{a}, \boldsymbol{b}\rangle \leq \|\boldsymbol{b}\|\|\boldsymbol{a}\|_* \leq \|\boldsymbol{b}\|^2/2 + \|\boldsymbol{a}\|_*^2/2$, specifically $\langle\eta\boldsymbol{g}^t, \left(\boldsymbol{x}^t - \boldsymbol{y}^{t+1}\right)\rangle \leq \frac{\alpha_\Phi}{2}\|\boldsymbol{y}^{t+1} - \boldsymbol{x}^t\|^2 + \frac{\eta^2}{2\alpha_\Phi}\|\boldsymbol{g}^t\|_*^2$.

Remark that, as mentioned in the paper, our approach follows the potential argument of Bansal et Gupta [4], which has been appeared recently in *Theory of Computing, pp. 1-32, vol 15, 2019*. In particular, the part related to the concern is proved in their paper (page 19, paragraph "Potential change", `https://theoryofcomputing.org/articles/v015a004/v015a004.pdf`). Note that they considered convex functions.

In conclusion, we believe that our proof is correct.