We thank Reviewers (R) 1, 2, 3, and 4 (who gave us marks 7, 6, 8, 6 respectively) for their positive feedback on the quality and clarity of the paper and their pertinent suggestions. About its writing, we are taking your remarks into account to improve the structure of some sections and we will correct the writing defaults. We will also add details on the concepts (e.g. R2-generalized geodesic convexity, and R3-why the proof differs from the Euclidean case) and discussion on the practical implementation. We first want to recall that our main contribution is the analysis of the FB scheme as an optimization algorithm to minimize $\mathcal{G} = \mathcal{E}_F + \mathcal{H}$. For instance, if $\mathcal{H}$ is the negative entropy, FB is: (i) faster than Langevin Monte Carlo (LMC), although we will acknowledge that it has an higher iteration complexity; (ii) unbiased (i.e. converges to $\mu^\star$ with a constant step) unlike LMC, because the F and B operators are adjoint to each other (see [34, Sec. F.1]). As R2 says, our work contrasts with many contributions on optimization over the Wasserstein space that only obtain guarantees in continuous time, which corresponds to an idealized setting.

R1, R2, R3, R4. Examples of functionals $\mathcal{H}$ in ML (other than entropies). We shall give formulas for the negative entropy and higher order entropies $\mathcal{H}$ that are natural regularizers. Besides, consider a Infinite-width 1 hidden layer neural network (1HLNN, see [3,12,23]) and let n be the number of neurons. For any input x, the output of the 1HLNN can be written as $f(x) = \frac{1}{n}\sum_{i=1}^n \phi(x, z_i)$, where $\phi(x, z_i) = w_i\psi(x, \theta_i)$ with $w_i \in \mathbb{R}$ the weight of the i-th neuron and $\theta_i \in \mathbb{R}^d$ parametrizes the activation function ψ (e.g. a sigmoid). Given (regression) data $(x, y) \sim p$, the optimization of this NN can be written as the minimization of the MSE: $\mathrm{argmin}_{z_1,\dots,z_n\in\mathcal{Z}} \frac{1}{2}\mathbb{E}_{(x,y)\sim p}[\|y - \frac{1}{n}\sum_{i=1}^n \phi(x, z_i)\|^2]$, where $\mathcal{Z} = \mathbb{R} \times \mathbb{R}^d$. When $n \to \infty$, this becomes an optimization problem over $\mathcal{P}(\mathcal{Z})$ (the set of probability measures over $\mathcal{Z}$) : $\mathrm{argmin}_{\mu\in\mathcal{P}(\mathcal{Z})} \frac{1}{2}\mathbb{E}_{(x,y)\sim p}[\|y - \int \phi(x, z)d\mu(z)\|^2]$. Expanding this loss function leads to the objective $\mathcal{G}(\mu) = \int F(z)d\mu(z) + \frac{1}{2}\int K(z, z')d\mu(z)d\mu(z')$, where $F(z) = -\mathbb{E}_{(x,y)\sim p}[y\phi(x, z)]$ is a potential and $K(z, z') = \mathbb{E}_{x\sim p_x}[\phi(x, z)\phi(x, z')]$ is an interaction term. Note however that this example is not strictly covered by our theory as $\mathcal{G}$ is only λ-geodesically convex (with $\lambda < 0$, see [12]).

R2, R4. Computation of JKOs. We agree and will acknowledge that the cases given by R1 and R2, as far as we know, are the only ones where exact formulas of JKOs are known. However, there is an extensive literature on the computation of JKOs of *generic* functionals using subroutines, see e.g. the review [29, Section 4.8] and the more recent work [27]. Just as proximal methods in Euclidean optimization, the FB scheme relies on subroutines to compute the JKO step. Our results on FB scheme could motivate research on specific JKOs relevant to ML. **Assumptions B1–B3.** B1-B2 are just general technical assumptions used in [2] (see [2, Eq 10.1.1a, Eq 10.1.1b]) that are always satisfied in relevant cases. Then [2, Prop 9.3.2, 9.3.5 and 9.3.9] gives broad examples of (potential, interaction and internal) energies satisfying B3, e.g., potential (resp. interaction) energies if the potential (resp. interaction) term is convex, and entropies.

R2, R3, R4. Experiments. We only provided simulations in a toy model, in low and high-dimensions, to illustrate some of our results such as the predicted linear convergence (Fig. 2). Further numerical investigations would imply integrating subroutines to tackle problems with more complicated JKOs, and will be the subject of future work. We will add documentation about the experiments, including the computer environment and the update formulas obtained from [34, App. G], and discuss more precisely our numerical results.

R1. Related work. In the following, [[]] denotes references mentioned in the reviews while the ones of the paper are still denoted by []. These splitting methods are indeed related, we will cite the missing references [[1,2,3,6]]. However, we stress that to the best of our knowledge, the FB scheme is new, except in the specific case where $\mathcal{H}$ is the negative entropy [34, App. G]. Also, it is not covered in [[1,2,3,6]]. The method of [[1]] employs the JKO of $\mathcal{E}_F + \mathcal{H}$ instead of splitting $\mathcal{E}_F$ and $\mathcal{H}$. Then, [[2]] and variants of the Langevin algorithm [[4,5,6]] are not FB because they use a flow step for $\mathcal{H}$, i.e. a step where a distribution is transported to another one following the *exact* gradient flow of $\mathcal{H}$. Indeed, in the Langevin algorithm, $\mathcal{H}$ is the negative entropy and the exact gradient flow is given by the Brownian motion (see [34, Sec. 2.2.2 vs Sec. 4.1] for the difference between FB and Langevin). Finally, [[3]] proposes an alternating scheme (Eq. (10) therein) which also differs from the FB scheme. **Non separable $\mathcal{E}_F$.** This is an interesting question. If $\mathcal{E}_F$ is replaced by some $\mathcal{F}$ not separable, i.e. $\mathcal{F}(\mu) \neq \int F(x)d\mu(x)$, then $\nabla\mathcal{F}(\mu) \neq \nabla F$ and Eq. (17) becomes $\nu_{n+1} = (I - \gamma\nabla\mathcal{F}(\mu_n))_{\#}\mu_n$. Extending our results is straightforward if $\mathcal{F}$ is convex along any interpolating curve (e.g. interaction energies with convex term [2, Prop. 9.3.5]) because one can still upper bound scalar products between $\nabla\mathcal{F}$ and *non-optimal* pushforwards (arising in the FB scheme). If $\mathcal{F}$ is only convex along geodesics, this is more difficult.

R2. **Related work.** We shall provide more details on [11,34]. [11] considers the resolution of the Wasserstein barycenter problem, which corresponds to a non geodesically convex objective, via gradient descent. This is done using nonconvex optimization techniques, but not using JKO steps. [34, App. G] inspired us the FB scheme. We extended it to a general $\mathcal{H}$ and established convergence rates (which were unknown even in the particular case of [34, App. G]).

R4. Contribution. We appreciate your positive comments on the inclusion of such papers to NeurIPS. We think that our contribution is not as incremental as it may look. The FB scheme was proposed in a particular case but was never analyzed as an optimization algorithm (even in the particular case, [34, App. G]). It is also surprising that the convergence rates are similar to the Hilbert case: LMC, which deals with $\mathcal{H}$ being the negative entropy, does not have this property. The proof requires non trivial ideas that are interesting on their own.