We thank to Reviewers 1, 2 and 3 (who gave us marks 7, 8 and 6, respectively) for their pertinent remarks.

R1+R3: Contribution. We agree that the optimization counterpart of SPLA can be related to Passty algorithm. However, it is a much more general optimization algorithm than Passty since all the functions are allowed to be expectations treated through stochastic gradients/stochastic prox. Therefore, we call it **Stochastic Passty**. In the optimization literature, the non asymptotic theory of this algorithm is still unknown. The only known particular cases are $n = 1$ and G_1 deterministic (prox-SGD), and the case $F = 0$ and $n = 1$ (Stochastic Proximal Point Algorithm, [29]). We were inspired by the proof structure of [17] (we will update this reference as it is published in JMLR), which is very adaptive. It allows to separate the analysis of SPLA into two pieces: the analysis of the optimization counterpart (here, Stochastic Passty) and the analysis of the Gaussian noise. Here, all the non asymptotic analysis of Stochastic Passty had to be done and involves modern tools of convex analysis such as random prox and random subdifferentials.

R1+R2+R3: Corollaries. We agree that we could provide more insights on the corollaries (Cor). As suggested by R2 and R3, we can compare the bounds with the one of [17]. First, in the particular case $n = 1$ and G_1 deterministic, SPLA boils down to the algorithm of [17, Section 4.2], Cor2 matches exactly Cor18 of [17] and Cor3 matches Cor22[1]. Cor4 has no counterpart in [17]. We now focus on the case $F = 0$ and $n = 1$ of SPLA, as it concentrates the innovations of our work. In this case, $L = 0$ and $\sigma_F = 0$. Compared to SSLA, our Cor2 matches with Cor14 of [17]. Actually, our constant C in Cor2 might be better because $C = L_{G_1}^2 \leq M^2 + D^2$ [1], due to the fact that we only need to bound the L2 norm of the minimal section (and not of any subgradient as in [17])[1]. In summary, [17] only covers the case $n = 1$ and G_1 deterministic of Cor2 and Cor3, and doesn't cover Cor4. The main advantage of SPLA over SSLA is its numerical stability (because SPLA is a proximal method [41], see the next paragraph).

R1+R2: Simulation. We agree that we could improve the experimental section by using a ground truth. We will add the following comparison of SSLA and SPLA in the case $F = 0$ and $n = 1$. Let $U = |x| = \mathbb{E}_\xi(|x| + x\xi)$ ($g_1(x, s) = |x| + xs$), where ξ is standard Gaussian. $\mu^\star \propto \exp(-U)$ is a standard Laplace distribution in $\mathbb{R}$. In this case, $L = \alpha = \sigma_F = 0$ and $C = L_{G_1}^2 = 2$. We shall illustrate the bound on $\mathrm{KL}(\mu_{\hat{x}^k} | \mu^\star)$ (Cor2 for SPLA and Cor14 of [17] for SSLA) for both algorithms using histograms. Note that the distribution $\mu_{\hat{x}^k}$ of $\hat{x}^k$ is a (deterministic) mixture of the μ_{x^j}: $\mu_{\hat{x}^k} = \frac{1}{k}\sum_{j=1}^k \mu_{x^j}$. Using Pinsker inequality, we can obtain a bound on the total variation distance between $\mu_{\hat{x}^k}$ and $\mu^\star$ from the bound on KL, and this can be illustrated by histograms[1]. In Figure 1, we take $\gamma = 10$ and do 10^5 iterations of both algorithms. Note that here the complexity of SPLA and SSLA are the same. SPLA enjoys the the well known advantages of proximal methods [41]: precision, numerical stability (less outliers), and robustness to step size.

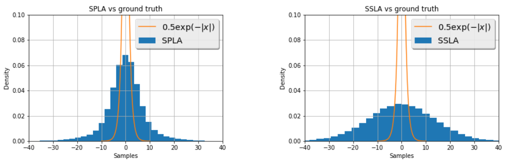

Figure 1: Comparison between histograms of SPLA and SSLA and the true density $0.5\exp(-|x|)$.

R2+R3: Motivations. There is an abundance of instances of the problem $\min U = \sum_{i=1}^n g_i$, where n is large, each $\mathrm{prox}_{\gamma g_i}$ has a closed form, but prox_U is intractable (as hard as minimizing U); e.g., SVM, logistic regression (see footnote Page 2), overlapping group lasso, TV regularization (see l. 210), see also [16, Section 2]. All these instances can be seen as MAP of $\propto \exp(-U)$ ([21,38,43,46]) and can be tackled by SPLA[1]. For the advantage of sampling a posteriori vs MAP for our example, see [19, Abstract, Paragraph 4.2.1 and 4.2.2]. Sampling allows to avoid overfitting[1].

R2: Minor comments. We especially thank R2 for his/her detailed comments. All minor comments will be easily addressed in the camera-ready version of the paper,[1] e.g., we will replace the sketch of the proof by a remark on gradient flows, remarks of R2 on l.382 and Lemma 6 are due to minor typos, and l. 449 and 477 will be easily clarified.

R3: Trade-offs. We shall illustrate our answer on l.212. As R3 suggested, n is analogous to the minibatch size in SGD. The larger n, the better the approximation of TV by the empirical mean (classical trade-off of SGD). Once n is fixed, one have to choose the level of splitting (*i.e* either treat the full sum in one prox or split each term of the sum). Less splitting is better: splitting is basically approximating the full prox by a combination of many prox (similar to the trade-off of SGD). As R3 says, we don't gain by splitting: one can check that the value of C doesn't change by treating g as $g/2 + g/2$, but in the latter case, two prox are needed at each iteration (so the computation time is twice). However, our key point in this work is that splitting is often unavoidable (see l. 210 and the paragraph "Motivations"). Finally, the value of C is smaller (better) if the noises impacting the g_i, $i \geq 2$ are independent[1].

[1] We will provide more details in the paper/supplementary but not here due to the lack of space.