Point-by-point reply to the referees' reports on NeurIPS2019

Paper ID : 6597
Paper title: Screening Sinkhorn Algorithm for Regularized Optimal Transport

We thank the Reviewers for their careful reports and for their positive comments. Below are our replies.

1 Answers to Reviewer #1

Question. *The method is not compared to Greenkhorn or any improved solver for this optimization problems.*

Answer.

Question. *The beauty of Sinkhorn is that it can be accelerated by GPU computations. It is unclear how the new algorithm would compete with a GPU-accelerated Sinkhorn algorithm.*

Answer.

Question. *The term "somewhat equivalent" at the end of the paper is not very scientific.*

Answer.

Question. *In Fig* 6 *even the smallest decimation leads to an immediate drop of performance. This suggests a fundamental and uncontroled difference in the solvers used. This observation is consistent with Figure* 7 *(low left on OTDA toy) where a decimation of* 2 *or less is slower to converge than a regular Sinkhorn.*

Answer.

Question. *There are quite weird results in Figure* 8 *too where a decimation of* 1.5 *or* 2 *leads to a drop in accuracy while the more strigent decimation factors do not incure any performance drop.*

Answer.

Question. *For figures when the legend and lines follow an order you should use a continuous paletter of colors and not categorical colors like here. It's then easier to see the gradient of speed gain with decimation.*

Answer.

Question. *Typos*

Answer. All the typos notified by the reviewer are corrected. We did also a whole rereading of the manuscript.

2 Answers to Reviewer #2

Thank you for taking time in reviewing our paper. Here are our point-by-point answers.

2.1 Main concerns

Question.

- *The first quantity in the bound, $C_{\max}/\eta$, is only small when η is large, which is the highly regularized regime. But in this regime, the Sinkhorn divergence is a poor estimate of OT. Indeed, it is known that η must be taken of size $O(\delta/\log n)$ for the Sinkhorn divergence to be within $\pm\delta$ of the OT value. But in this regime, the $C_{\max}/\eta$ term in your bound scales in the accuracy δ extremely poorly as $1/\delta$.*

- *For example, in practice, η is typically taken to be around $1/10$ or $1/20$ (if not smaller). Even in this so-called "moderately regularized regime", $C_{\max}/\eta$ is extremely large, e.g. $10{*}C_{\max}$ or $20{*}C_{\max}\dots$ (Of course there is the hidden constant in the big-O in your bound. It would be helpful to know what this constant is, but I suspect it is not much smaller than 1, if at all. This would be unfortunate since approximating OT within $C_{\max}$ is trivial, and can be done without even looking at the data: simply output the independent coupling $\mu * \nu^\top$.)*

- *To summarize, in the regularization regime where the Sinkhorn divergence actually approximates OT well (i.e. $\delta \ll 1$ small, and $\eta \approx \delta$), the error in the presented bound appears to be so big (namely $C_{\max}/\eta$), to the extent that it is perhaps bigger than $C_{\max}$ (which would then be a meaningless bound, see above comment). Am I missing something?*

Answer.

Question.

- *I am concerned that the terms $||\mu - \mu^{sc}||_1$ and $||\nu - \nu^{sc}||_1$ in the bound, can be large. My concern arises from the fact that for matrix scaling (which is exactly the dual Sinkhorn divergence), there are matrices for which the dual optimal solutions u and v are s.t. the range of e^{u_i} and e^{v_j} is exponentially large in n. See e.g., page 10 of the paper.*

- *Kalantari-Khachiyan 1996 "On the complexity of non-negative matrix scaling" (For further discussion of such issues, see also e.g. Cohen et al. 2017 "Matrix scaling and balancing...")*

- *From what I remember, these hard examples also apply for 'approximate' scaling, i.e. whenever u, v do not have exponentially large ranges, then $diag(e^u)K\, diag(e^v)$ has marginals which are very far from the desired marginals μ, ν. In other words, for such inputs, $||\mu - \mu^{sc}||_1$ and $||\nu - \nu^{sc}||_1$ can only be small if the thresholds in (4) are exponentially small in n. But this requires taking your screening parameter*

> *eps to be exponentially small in n, which means that $c_{\mu,\nu}$ will also be exponentially small in n, yielding a dependence of $\log(1/c_{\mu,\nu}) = poly(n)$ in the Proposition 3 bound, which is an extremely massive error bound. Am I missing something?*

Answer.

Question. *It is unsatisfactory that ω_κ in the bound is not defined in the proposition statement, and is only stated to decay as $o(1)$ as the screening parameter κ tends to 1. It is critical to understand how fast this error term decays in terms of this parameter choice, since the point of this proposed screening approach is to trade off runtime (i.e. screen more aggressively, partially done by taking κ *far from* 1) with accuracy (i.e. ensure screening does not change the value of the Sinkhorn divergence much, partially done by taking κ *close* to 1).*

Answer.

2.2 Other comments

Question. *The proposed approach is to solve the screened (i.e. smaller-size) Sinkhorn divergence problem via L-BGFS. Why not use the standard Sinkhorn alternating minimization typically used for Sinkhorn divergences? Changing two things (screening for pre-processing, and L-BFGS for solving) makes it difficult for the reader to understand the affect of each change.*

Answer.

Question. *Something seems a bit suspect about the proposed formulation (3), in that (unless you set the parameter $\kappa = 1$, which you purposely do not do) this optimization problem is not invariant under adding a constant times the all-ones vector to u, and subtracting the same from v. This is an important feature of matrix scaling / standard Sinkhorn divergences.*

Answer.

Question. *Section 3: the plots are not reproducible (or as informative as they should be) since it is not stated what the distribution of x_i and y_j are/look like.*

Answer.

Question. *L45: I suggest changing "reformulation" to a "new formulation", as the proposed formulation is not the same as Sinkhorn divergences, but rather is an altered formulation of it. (This is written a few times in the paper; the same comment applies throughout.)*

Answer.

Question. *I believe the term $\log(1/K_{\min}^2)$ in the bound can be removed, as it is equal to $\log(1/e^{-2*C_{\max}/\eta}) = 2C_{\max}/\eta$, and thus absorbed by the first term in the bound.*

Answer.

Question. *L64: that is entropy, not negative entropy*

Answer.

Question. *When describing the preliminaries in Section 2, it should be briefly mentioned that everything there is expository.*

Answer.

Question. *End of first paragraph: technically, the LP solver in [Lee-Sidford 2014] can solve OT exactly in $\tilde{O}(n^{2.5})$ time. Granted, there is currently no practical implementation of this complicated algorithm, but it should be mentioned there while you are describing runtimes of existing methods.*

Answer.

2.3 Comments on exposition

Question. *Typos:*
— L31-32: sentence should be re-written
— L103-104: problems with tense
- References: missing capitalization in some (but not all) titles, e.g. pca, wasserstein, lagrangian.

Answer.

3 Answers to Reviewer #3

Question. *The article is interesting and participates to a growing literature on fast approximations of optimal transport. The experiments are quite convincing: the proposed method can provide a way of reducing the cost while maintaining a reasonable accuracy. The gains are not always groundbreaking but clearly significant and the method appears to be widely applicable. It is also very nice that users can specify the approximation parameters in terms of n_b and m_b, rather than ε and κ.*

Answer.

Question. *One criticism is that the method is compared only with the "vanilla" Sinkhorn algorithm, so it is not easy to see whether the proposed method would provide gains compared to other recently proposed variants, such as the Greenkhorn algorithm mentioned in the introduction. It is also unclear whether the proposed thresholding strategy could be combined with these alternative techniques, e.g. with Greenkhorn, for further gains, or whether it is incompatible.*

Answer.

Question. *In terms of presentation, it would be helpful to have a self-contained pseudocode, that would not rely on a routine such as "L-BFGS-B", although that can be considered to be quite standard. I also wonder whether L-BFGS-B is the only algorithm applicable for step 2 of Algorithm 1, or if there are other sensible alternatives there; L-BFGS-B is presented as the only option.*

Answer.

Question. *Typos*

Answer. All the typos notified by the reviewer are corrected. We did also a whole rereading of the manuscript.

p8: "features have been extracted from the feature extractor" is awkward.