Online Learning for Latent Dirichlet Allocation: Supplementary Material

Anonymous Author(s)
Affiliation
Address
email

1 Analysis of online VB for LDA with randomized E step

In the analysis of online VB for LDA with a deterministic (but approximate) E step, we used $g(n)$ to denote a population distribution over documents' word counts n defined by a corpus:

$$g(n) \triangleq \frac{1}{D} \sum_{d=1}^{D} \mathbb{I}[n = n_d]. \quad (1)$$

To analyze the case where a randomized E step is used to fit the per-document variational parameters γ and ϕ to a document d and the variational topic parameters $\boldsymbol{\lambda}$, we redefine g as a joint distribution over both documents and per-document parameters:

$$g(\gamma, \phi, n|\boldsymbol{\lambda}) = g(\gamma, \phi|n, \boldsymbol{\lambda})g(n) = \frac{1}{D} \sum_{d=1}^{D} \mathbb{I}[n = n_d]g(\gamma, \phi|n, \boldsymbol{\lambda}). \quad (2)$$

$g(\gamma, \phi|n, \boldsymbol{\lambda})$ is the distribution over per-document variational parameters γ and ϕ implicitly defined by the randomized algorithm for fitting γ and ϕ given a document's word count vector n and the top-level variational parameters $\boldsymbol{\lambda}$. $g(n)$ is the population distribution, as before. We sample from $g(\gamma, \phi, n|\boldsymbol{\lambda})$ by choosing a document d at random from the corpus, and then using a randomized E step to fit γ_d and ϕ_d given n_d and $\boldsymbol{\lambda}$.

Our goal is now to find a setting of $\boldsymbol{\lambda}$ that optimizes the objective

$$\mathcal{L}(g, \boldsymbol{\lambda}) \triangleq D\mathbb{E}_g[\ell(n, \gamma, \phi, \boldsymbol{\lambda})|\boldsymbol{\lambda}]. \quad (3)$$

Optimizing this objective means finding a setting of $\boldsymbol{\lambda}$ that gives us as high an *expected* value of the ELBO $\mathcal{L}$ as possible after running our randomized E step. Deterministic optimization of this objective is impossible. We cannot assume an analytic form for $g(\gamma, \phi|n, \boldsymbol{\lambda})$, and so we cannot even compute this objective. We can nonetheless optimize it using stochastic gradient.

If $n_t, \gamma_t, \phi_t \sim g$, then

$$\mathbb{E}_g[\nabla_\lambda D\ell(n_t, \gamma_t, \phi_t, \boldsymbol{\lambda})|\boldsymbol{\lambda}] = \nabla_\lambda D\mathbb{E}_g[\ell(n, \gamma, \phi, \boldsymbol{\lambda})|\boldsymbol{\lambda}] = \nabla_\lambda \mathcal{L}(g, \boldsymbol{\lambda}). \quad (4)$$

That is, the expected value of the gradient of the per-document objective is equal to the gradient of the objective $\mathcal{L}$, satisfying a critical condition for convergence [1]. This differs from the usual stochastic gradient setup in that we cannot directly compute the objective we are optimizing, but are instead trying to find a stationary point of an *expected* objective. Since $\ell(n, \gamma, \phi, \boldsymbol{\lambda})$ is thrice differentiable with respect to $\boldsymbol{\lambda}$, its expectation is as well, satisfying the first assumption in section 5 of [1]. (This is true even though we cannot compute the expectation or its derivatives.) It can be easily verified that the other conditions for general online optimization outlined in [1] are satisfied.

2 Supplemental learning parameter evaluation figures

Below are three plots summarizing the held-out perplexities obtained by our online VB algorithm on sets of Wikipedia and *Nature* articles for every setting of the learning parameters κ (learning rate), τ_0 (downweighting of the initial iterations) and S (mini-batch size) that we evaluated.

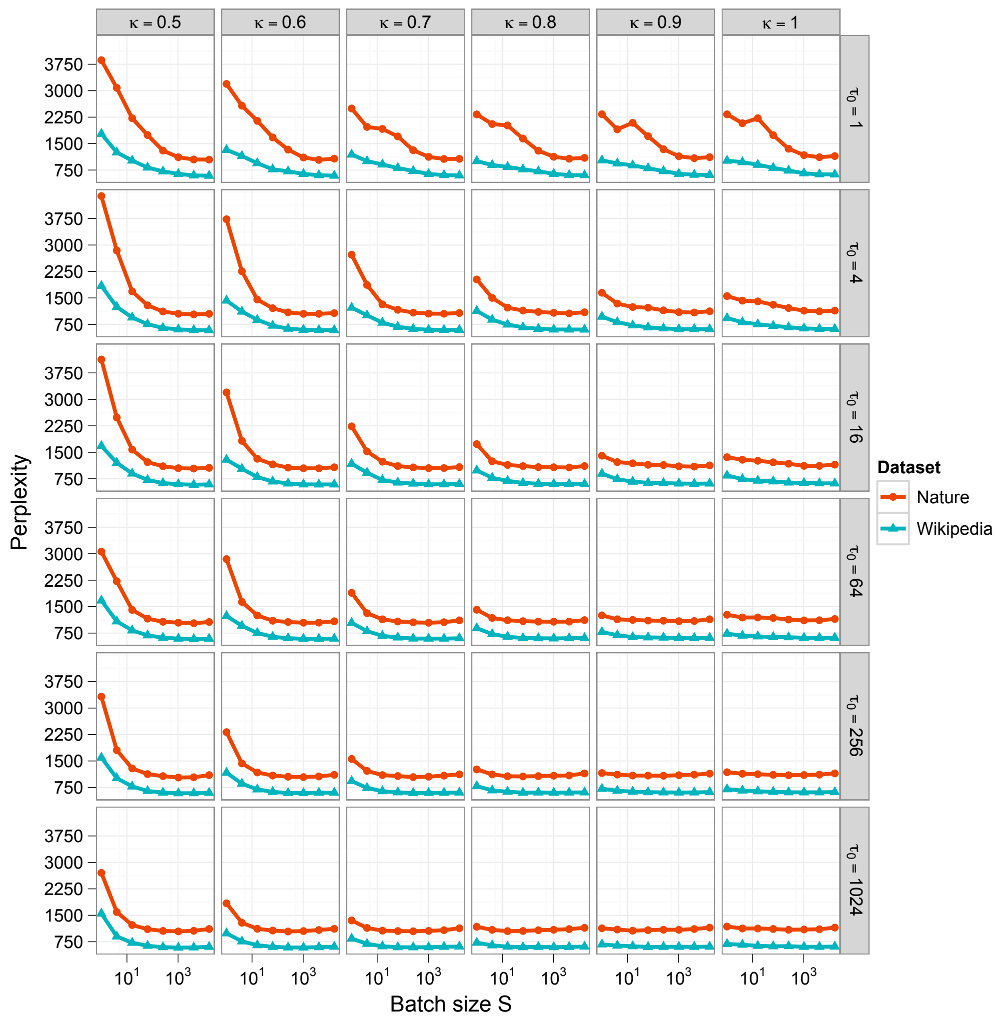

Figure 1: Held-out perplexity obtained on the *Nature* and Wikipedia corpora for various settings of the learning rate κ, mini-batch size S, and initial slowdown parameter τ_0, presented as a function of mini-batch size S.

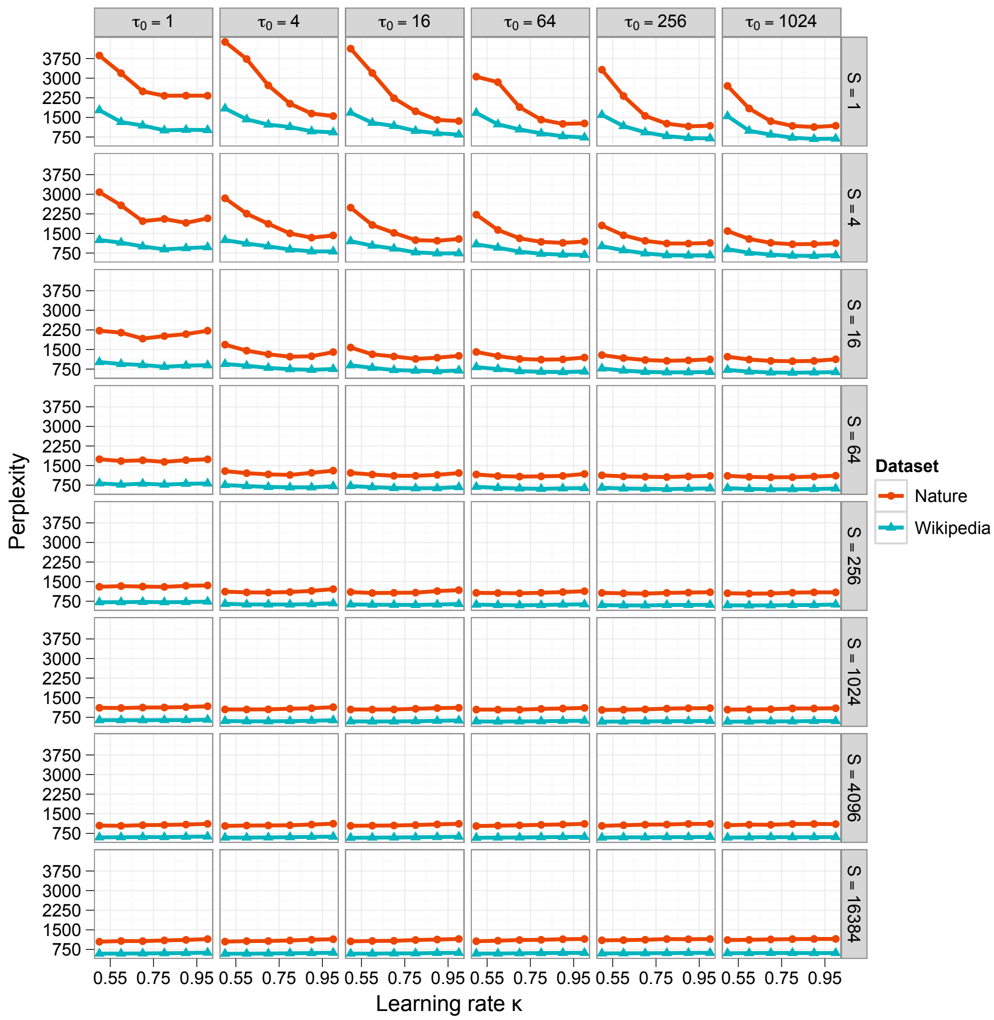

Figure 2: Held-out perplexity obtained on the *Nature* and Wikipedia corpora for various settings of the learning rate κ, mini-batch size S, and initial slowdown parameter τ_0, presented as a function of learning rate κ.

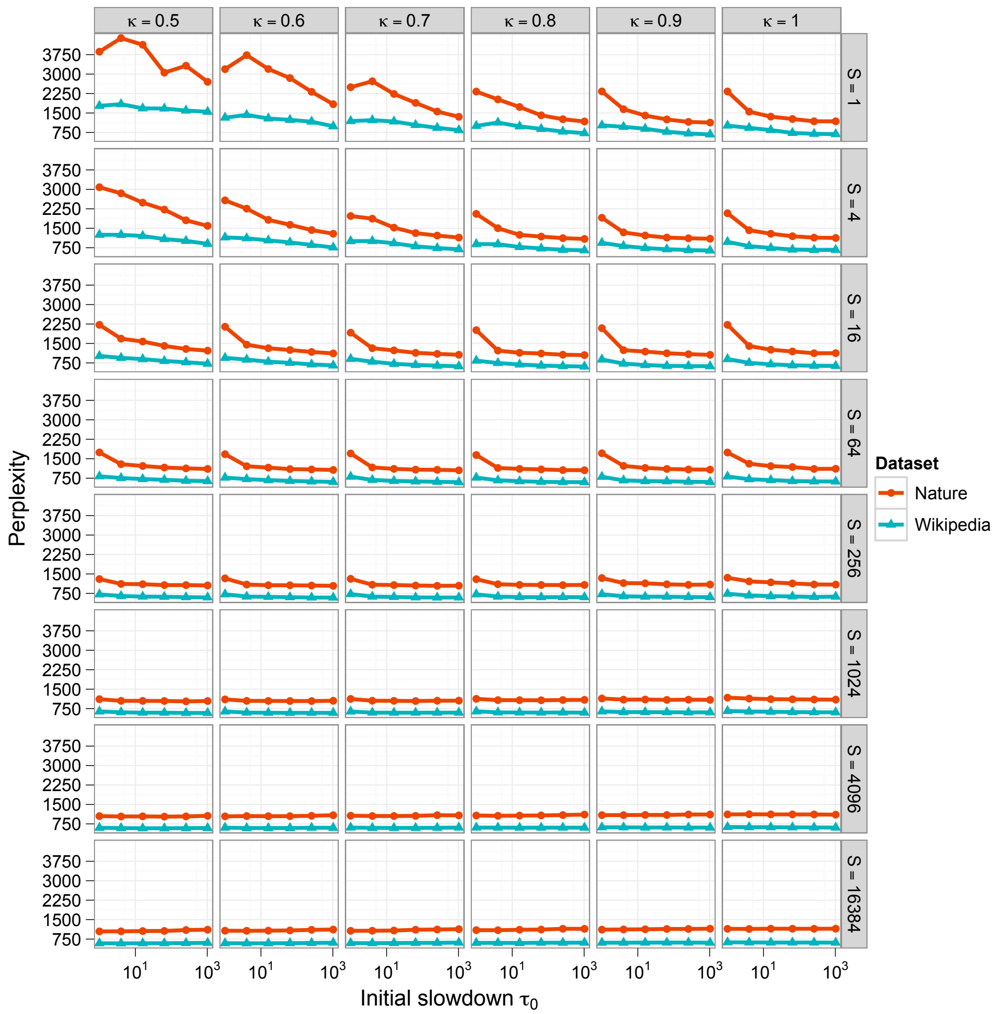

Figure 3: Held-out perplexity obtained on the *Nature* and Wikipedia corpora for various settings of the learning rate κ, mini-batch size S, and initial slowdown parameter τ_0, presented as a function of initial slowdown τ_0.

References

[1] L. Bottou. *Online learning and stochastic approximations*. Cambridge University Press, Cambridge, UK, 1998.