

000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053

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 λ , we redefine g as a joint distribution over both documents and per-document parameters:

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

$g(\gamma, \phi | n, \lambda)$ 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 λ . $g(n)$ is the population distribution, as before. We sample from $g(\gamma, \phi, n | \lambda)$ 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 λ .

Our goal is now to find a setting of λ that optimizes the objective

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

Optimizing this objective means finding a setting of λ 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 $g(\gamma, \phi | n, \lambda)$, 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

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

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 $\ell(n, \gamma, \phi, \lambda)$ is thrice differentiable with respect to λ , 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.

054
 055
 056
 057
 058
 059
 060
 061
 062
 063
 064
 065
 066
 067
 068
 069
 070
 071
 072
 073
 074
 075
 076
 077
 078
 079
 080
 081
 082
 083
 084
 085
 086
 087
 088
 089
 090
 091
 092
 093
 094
 095
 096
 097
 098
 099
 100
 101
 102
 103
 104
 105
 106
 107

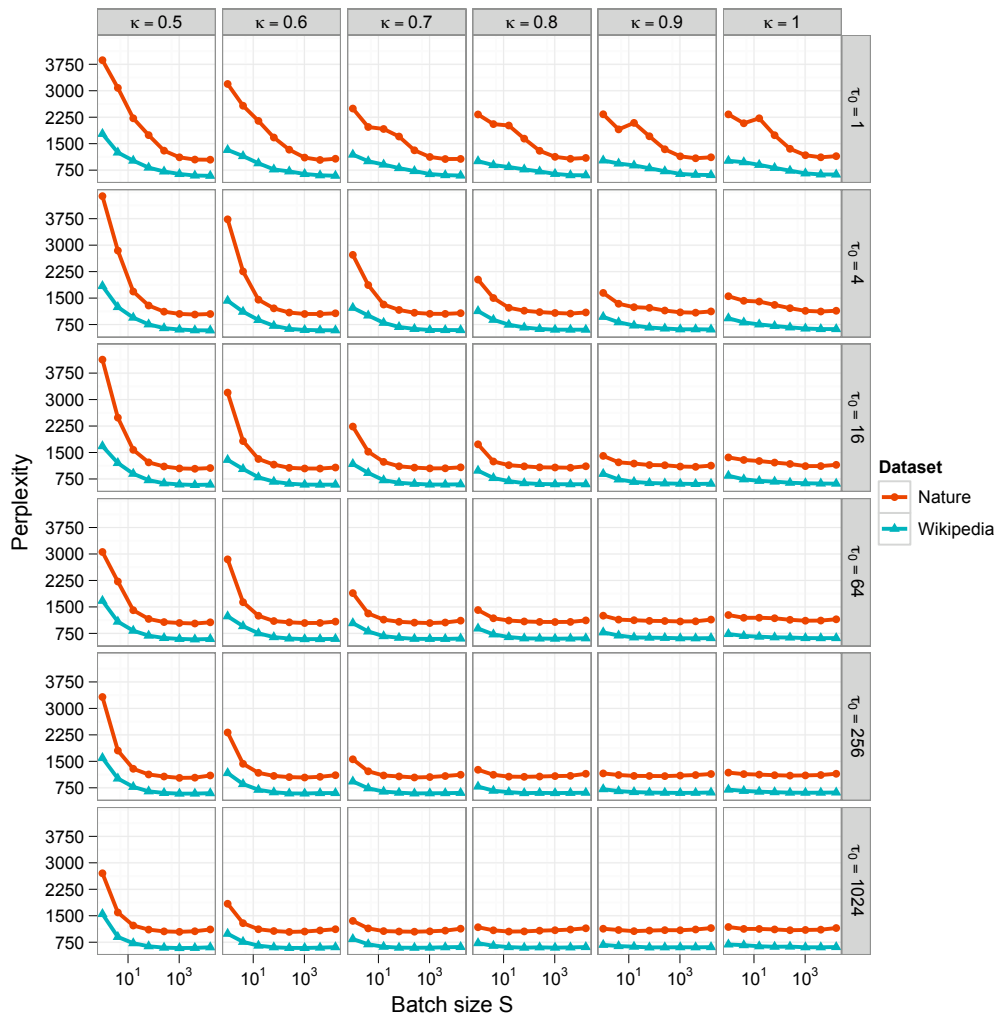


Figure 1: Held-out perplexity obtained on the *Nature* and Wikipedia corpora for various settings of the learning rate κ , mini-batch size S , and initial slowdown parameter τ_0 , presented as a function of mini-batch size S .

108
 109
 110
 111
 112
 113
 114
 115
 116
 117
 118
 119
 120
 121
 122
 123
 124
 125
 126
 127
 128
 129
 130
 131
 132
 133
 134
 135
 136
 137
 138
 139
 140
 141
 142
 143
 144
 145
 146
 147
 148
 149
 150
 151
 152
 153
 154
 155
 156
 157
 158
 159
 160
 161

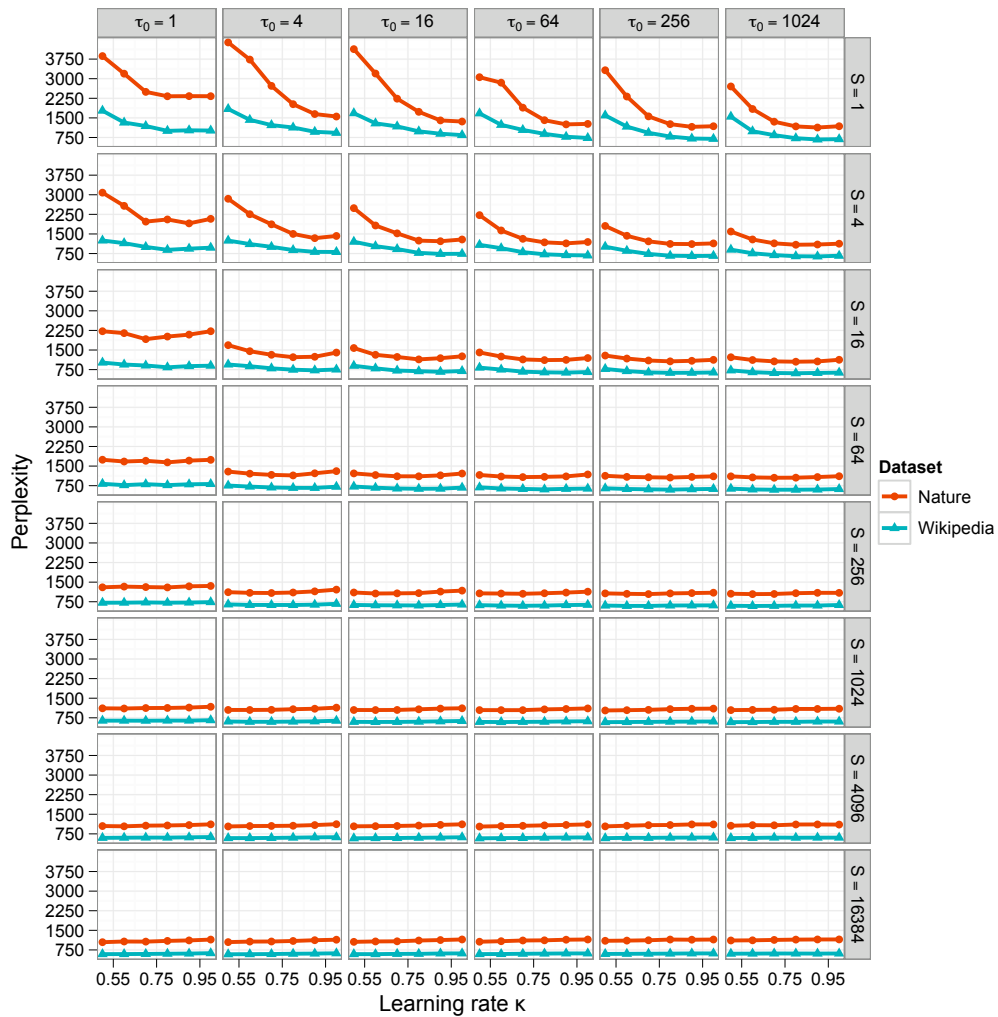


Figure 2: Held-out perplexity obtained on the *Nature* and *Wikipedia* corpora for various settings of the learning rate κ , mini-batch size S , and initial slowdown parameter τ_0 , presented as a function of learning rate κ .

162
 163
 164
 165
 166
 167
 168
 169
 170
 171
 172
 173
 174
 175
 176
 177
 178
 179
 180
 181
 182
 183
 184
 185
 186
 187
 188
 189
 190
 191
 192
 193
 194
 195
 196
 197
 198
 199
 200
 201
 202
 203
 204
 205
 206
 207
 208
 209
 210
 211
 212
 213
 214
 215

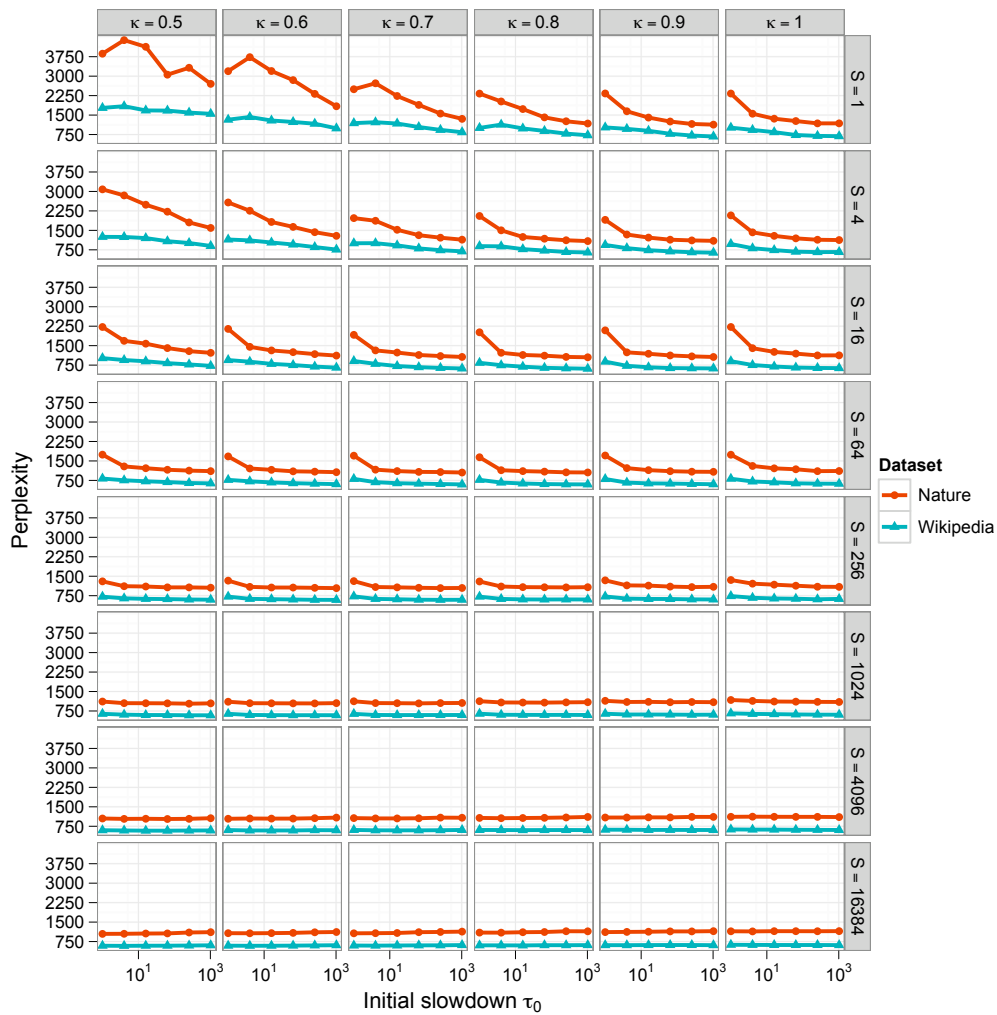


Figure 3: Held-out perplexity obtained on the *Nature* and Wikipedia corpora for various settings of the learning rate κ , mini-batch size S , and initial slowdown parameter τ_0 , presented as a function of initial slowdown τ_0 .

216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269

References

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