SUPPLEMENTARY MATERIAL
Optimal Mini-Batch and Step Sizes for SAGA

A. Proofs of the Upper Bounds of $\mathcal{L}$

A.1. Master lemma

Proof of Lemma 1. Since the f_i's are convex, each realization of f_v is convex, and it follows from equation 2.1.7 in (Nesterov, 2014) that

$$\|\nabla f_v(x) - \nabla f_v(y)\|_2^2 \quad \leq \quad 2L_v \left(f_v(x) - f_v(y) - \langle \nabla f_v(y), x - y\rangle\right). \tag{24}$$

Taking expectation over the sampling gives

$$\begin{aligned}
\mathbb{E}[\|\nabla f_v(x) - \nabla f_v(x^*)\|_2^2] &\leq 2\mathbb{E}\left[L_v \left(f_v(x) - f_v(x^*) - \langle \nabla f_v(x^*), x - x^*\rangle\right)\right] \\
&\overset{(24)}{=} \frac{2}{n}\mathbb{E}\left[\sum_{i=1}^n L_v v_i \left(f_i(x) - f_i(x^*) - \langle \nabla f_i(x^*), x - x^*\rangle\right)\right] \\
&= \frac{2}{n}\sum_{i=1}^n \mathbb{E}\left[L_v v_i\right]\left(f_i(x) - f_i(y) - \langle \nabla f_i(x^*), x - x^*\rangle\right) \\
&\leq 2 \max_{i=1,\ldots,n} \mathbb{E}\left[L_v v_i\right]\left(f(x) - f(x^*) - \langle \nabla f(x^*), x - x^*\rangle\right) \\
&= 2 \max_{i=1,\ldots,n} \mathbb{E}\left[L_v v_i\right]\left(f(x) - f(x^*)\right).
\end{aligned}$$

where in the last equality the full gradient vanishes because it is computed at optimality. The result now follows by comparing the above with the definition of expected smoothness in (7). □

A.2. Proof of the simple bound

Proof of Theorem 2. To derive this bound on $\mathcal{L}$ we use that

$$L_B \leq \frac{1}{b}\sum_{j\in B} L_j \ , \tag{25}$$

which follows from repeatedly applying Lemma 8. For $b \geq 2$, it follows from Equation (17) and Equation (25) that

$$\mathcal{L} \leq \frac{1}{b\binom{n-1}{b-1}} \max_{i=1,\ldots,n} \left\{ \sum_{\substack{B\subseteq[n]:\\ |B|=b\wedge i\in B}} \sum_{j\in B} L_j \right\}. \tag{26}$$

Using a double counting argument we can show that

$$\begin{aligned}
\sum_{\substack{B\subseteq[n]:\\ |B|=b\wedge i\in B}} \sum_{j\in B} L_j &= \sum_{j=1}^n \sum_{\substack{B\subseteq[n]:\\ |B|=b\wedge i,j\in B}} L_j \\
&= \sum_{j\neq i} \sum_{\substack{B\subseteq[n]:\\ |B|=b\wedge i,j\in B}} L_j + \sum_{\substack{B\subseteq[n]:\\ |B|=b\wedge i\in B}} L_i \\
&= \sum_{j\neq i} \binom{n-2}{b-2} L_j + \binom{n-1}{b-1} L_i \\
&= \binom{n-2}{b-2}(n\bar{L} - L_i) + \binom{n-1}{b-1} L_i \ .
\end{aligned} \tag{27}$$

Inserting this into Equation (26) gives

$$\begin{aligned}\mathcal{L} &\leq \frac{1}{b\binom{n-1}{b-1}} \max_{i=1,\dots,n} \left\{ \binom{n-2}{b-2} n\bar{L} + \left(\binom{n-1}{b-1} - \binom{n-2}{b-2} \right) L_{\max} \right\} \\ &= \frac{n\binom{n-2}{b-2}}{b\binom{n-1}{b-1}} \bar{L} + \frac{\binom{n-1}{b-1} - \binom{n-2}{b-2}}{b\binom{n-1}{b-1}} L_{\max} \\ &= \frac{n}{b}\frac{b-1}{n-1}\bar{L} + \frac{1}{b}\frac{n-b}{n-1} L_{\max} \ . \end{aligned} \tag{28}$$

We also verify that this bound is valid for 1-nice sampling. Indeed, we already have that in this case $\mathcal{L} = L_{\max}$. □

A.3. Proof of the Bernstein bound

To start the proof of Theorem 3, we re-write the expected smoothness constant as the maximum over an expectation. Let S^i be a $(b-1)$-nice sampling over $[n] \setminus \{i\}$. We can write

$$\begin{aligned}\mathcal{L} &= \frac{1}{\binom{n-1}{b-1}} \max_{i=1,\dots,n} \left\{ \sum_{\substack{B \subseteq [n]: \\ |B|=b \wedge i \in B}} L_B \right\} \\ &= \max_{i=1,\dots,n} \mathbb{E}\left[L_{S^i \cup \{i\}} \right] \\ &\overset{\text{Lemma 2}}{=} \max_{i=1,\dots,n} U\mathbb{E}\left[\lambda_{\max}\left(\frac{1}{b} \sum_{j \in S^i \cup \{i\}} a_j a_j^\top \right) \right] \ . \end{aligned} \tag{29}$$

One can come back to the definition of the subsample smoothness constant Equation (12) and interpret previous expression as an expectation of the largest eigenvalue of a sum of matrices. This insight allows us to apply a matrix Bernstein inequality, see Theorem 7, to bound $\mathcal{L}$.

For the proof of Theorem 3, we first need the two following results.

Lemma 4. *Let $a_j \in \mathbb{R}^d$, $i \in \{1, \dots, n\}$ and let S^i be a $(b-1)$-nice sampling over the set $[n] \setminus \{i\}$. It follows that*

$$\mathbb{E}\left[\sum_{j \in S^i \cup \{i\}} a_j a_j^\top \right] = a_i a_i^\top + \frac{b-1}{n-1} \sum_{j=1, j \neq i}^{n} a_j a_j^\top \ . \tag{30}$$

Proof of Lemma 4. This results follows using a double-counting argument at the fourth line of the computation.

$$\begin{aligned}
\mathbb{E}\left[\sum_{j\in S^i\cup\{i\}} a_j a_j^\top\right] &= \frac{1}{\binom{n-1}{b-1}} \sum_{\substack{B\subseteq[n]\setminus\{i\}:\\ |B|=b-1}} \sum_{j\in B\cup\{i\}} a_j a_j^\top \\
&= \frac{1}{\binom{n-1}{b-1}} \left(\binom{n-1}{b-1} a_i a_i^\top + \sum_{\substack{B\subseteq[n]\setminus\{i\}:\\ |B|=b-1}} \sum_{j\in B} a_j a_j^\top \right) \\
&= a_i a_i^\top + \frac{1}{\binom{n-1}{b-1}} \sum_{\substack{B\subseteq[n]\setminus\{i\}:\\ |B|=b-1}} \sum_{j\in B} a_j a_j^\top \\
&= a_i a_i^\top + \frac{1}{\binom{n-1}{b-1}} \sum_{j=1, j\neq i}^{n} \sum_{\substack{B\subseteq[n]\setminus\{i\}:\\ |B|=b-1\wedge j\in B}} a_j a_j^\top \\
&= a_i a_i^\top + \frac{\binom{n-2}{b-2}}{\binom{n-1}{b-1}} \sum_{j=1, j\neq i}^{n} a_j a_j^\top \\
&= a_i a_i^\top + \frac{b-1}{n-1} \sum_{j=1, j\neq i}^{n} a_j a_j^\top \ .
\end{aligned}$$

□

We then introduce another two lemmas which give a first intermediate bound.

Lemma 5. *Let $a_j \in \mathbb{R}^d$ for $j \in [n]$, let $i \in [n]$ and let S^i be a $(b-1)$-nice sampling over $[n]\setminus\{i\}$. We have*

$$\begin{aligned}
U\mathbb{E}\left[\lambda_{\max}\left(\frac{1}{b}\sum_{j\in S^i\cup\{i\}} a_j a_j^\top\right)\right] \leq & \frac{1}{b(n-1)}\left((n-b)L_i + n(b-1)L\right) \\
& + U\mathbb{E}\left[\lambda_{\max}\left(\frac{1}{b}\sum_{j\in S^i} a_j a_j^\top - \frac{1}{b}\frac{b-1}{n-1}\sum_{j\in[n]\setminus\{i\}} a_j a_j^\top\right)\right] \ . \qquad (31)
\end{aligned}$$

Proof of Lemma 5. Expanding the expectation we have

$$\begin{aligned}
&\mathbb{E}\left[\lambda_{\max}\left(\frac{1}{b}\sum_{j\in S^i\cup\{i\}} a_j a_j^\top\right)\right] \\
&\leq \lambda_{\max}\left(\mathbb{E}\left[\frac{1}{b}\sum_{j\in S^i\cup\{i\}} a_j a_j^\top\right]\right) + \mathbb{E}\left[\lambda_{\max}\left(\frac{1}{b}\sum_{j\in S^i\cup\{i\}} a_j a_j^\top - \mathbb{E}\left[\sum_{j\in S^i\cup\{i\}} a_j a_j^\top\right]\right)\right] \\
&= \frac{1}{b}\lambda_{\max}\left(a_i a_i^\top + \frac{b-1}{n-1}\sum_{j=1,\, j\neq i}^{n} a_j a_j^\top\right) + \mathbb{E}\left[\lambda_{\max}\left(\frac{1}{b}\sum_{j\in S^i\cup\{i\}} a_j a_j^\top - \frac{1}{b}\left(a_i a_i^\top + \frac{b-1}{n-1}\sum_{j=1,\, j\neq i}^{n} a_j a_j^\top\right)\right)\right] \\
&= \frac{1}{b}\lambda_{\max}\left(\frac{1}{n-1}\left((n-b)a_i a_i^\top + (b-1)\sum_{j=1}^{n} a_j a_j^\top\right)\right) + \mathbb{E}\left[\lambda_{\max}\left(\frac{1}{b}\sum_{j\in S^i} a_j a_j^\top - \frac{1}{b}\frac{b-1}{n-1}\sum_{j\in[n]\setminus\{i\}} a_j a_j^\top\right)\right] \\
&\leq \frac{1}{b(n-1)}\left((n-b)\frac{L_i}{U} + n(b-1)\frac{L}{U}\right) + \mathbb{E}\left[\lambda_{\max}\left(\frac{1}{b}\sum_{j\in S^i} a_j a_j^\top - \frac{1}{b}\frac{b-1}{n-1}\sum_{j\in[n]\setminus\{i\}} a_j a_j^\top\right)\right] \ ,
\end{aligned}$$

where in the first inequality we add and remove the mean and then apply Lemma 8. In the second equality we explicit the mean with Lemma 4 and in the last inequality we use again Lemma 8 for the left-hand side term. Finally, we multiply by U on both sides of the inequality. □

We recall the following lemma used to introduced the *practical estimate* given by

$$\mathcal{L}_{\text{practical}}(b) \overset{\text{Definition 5}}{:=} \frac{n}{b}\frac{b-1}{n-1}L + \frac{1}{b}\frac{n-b}{n-1}L_{\max} \ .$$

Lemma 3. *Let $a_j \in \mathbb{R}^d$ for $j \in [n]$ and let S^i be a $(b-1)$-nice sampling over $[n] \setminus \{i\}$, for every $i \in [n]$. It follows that*

$$\mathcal{L} \leq \mathcal{L}_{\textit{practical}}(b) + U \max_{i\in[n]} \mathbb{E}\left[\lambda_{\max}\left(N_i\right)\right] \ , \tag{21}$$

with $N_i := \frac{1}{b}\sum_{j\in S^i} a_j a_j^\top - \frac{1}{b}\frac{b-1}{n-1}\sum_{j\in[n]\setminus\{i\}} a_j a_j^\top$.

Proof of Lemma 3. The result comes from applying re-writing $\mathcal{L}$ as an expectation of the largest eigenvalue of a sum of matrices. Then we apply Lemma 5 and then taking the maximum over all $i \in [n]$. Thus, we have

$$\begin{aligned}
\mathcal{L} &\overset{(29)}{=} \max_{i=1,\dots,n} U\mathbb{E}\left[\lambda_{\max}\left(\frac{1}{b}\sum_{j\in S^i\cup\{i\}} a_j a_j^\top\right)\right] \\
&\overset{\text{Lemma 5}}{\leq} \frac{1}{b(n-1)}\left((n-b)L_i + n(b-1)L\right) + U\mathbb{E}\left[\lambda_{\max}\left(\frac{1}{b}\sum_{j\in S^i} a_j a_j^\top - \frac{1}{b}\frac{b-1}{n-1}\sum_{j\in[n]\setminus\{i\}} a_j a_j^\top\right)\right] \\
&\leq \frac{n}{b}\frac{b-1}{n-1}L + \frac{1}{b}\frac{n-b}{n-1}L_{\max} + \max_{i=1,\dots,n} U\mathbb{E}\left[\lambda_{\max}\left(\frac{1}{b}\sum_{j\in S^i} a_j a_j^\top - \frac{1}{b}\frac{b-1}{n-1}\sum_{j\in[n]\setminus\{i\}} a_j a_j^\top\right)\right] \ .
\end{aligned}$$

□

Proof of Theorem 3. Applying the previous lemma we get

$$\mathcal{L} \overset{(21)}{\leq} \frac{n}{b}\frac{b-1}{n-1}L + \frac{1}{b}\frac{n-b}{n-1}L_{\max} + \max_{i=1,\dots,n} U\mathbb{E}\left[\lambda_{\max}\left(N\right)\right] \ , \tag{32}$$

with $N := \frac{1}{b}\sum_{j\in S^i} a_j a_j^\top - \frac{1}{b}\frac{b-1}{n-1}\sum_{j\in[n]\setminus\{i\}} a_j a_j^\top$.

To further our argument, we will encode different samplings using unit coordinate vectors. Let $e_1, \dots, e_n \in \mathbb{R}^n$ be the unit coordinate vectors. Let $S^i = \{S_1^i, \dots, S_b^i\}$ denote an arbitrary but fixed ordering of the elements of S^i. With this we can encode the sampling without replacement as

$$\sum_{j\in S^i} a_j a_j^\top = \sum_{k=1}^{b-1}\sum_{j\in[n]\setminus\{i\}} (e_j)_{S_k^i} a_j a_j^\top \ . \tag{33}$$

Using this notation, the matrix N which can be further decomposed as

$$\begin{aligned}
N &= \frac{1}{b}\sum_{j\in S^i} a_j a_j^\top - \frac{1}{b}\frac{b-1}{n-1}\sum_{j\in[n]\setminus\{i\}} a_j a_j^\top \\
&= \frac{1}{b}\sum_{k=1}^{b-1}\sum_{j\in[n]\setminus\{i\}} (e_j)_{S_k^i} a_j a_j^\top - \frac{1}{b}\frac{b-1}{n-1}\sum_{j\in[n]\setminus\{i\}} a_j a_j^\top \\
&= \sum_{k=1}^{b-1}\frac{1}{b}\sum_{j\in[n]\setminus\{i\}}\left((e_j)_{S_k^i} - \frac{1}{n-1}\right) a_j a_j^\top \\
&:= \sum_{k=1}^{b-1} M_k \ .
\end{aligned}$$

where we have encoded the sampling S^i using unit coordinate vectors. The matrices $M_1, \ldots, M_{b-1}$ are sampled *without* replacement from the set

$$\left\{ \sum_{j \in [n]\setminus\{i\}} \frac{1}{b}\left(x_j - \frac{1}{n-1}\right) a_j a_j^\top \ : \ x \in \{e_1, \ldots, e_{i-1}, e_{i+1}, \ldots, e_n\} \right\} \ . \tag{34}$$

Now let $X_1, \ldots, X_b$ be matrices sampled *with* replacement from (34) and let $X_k := \frac{1}{b}\sum_{j\in[n]\setminus\{i\}} \left(z_j^k - \frac{1}{n-1}\right) a_j a_j^\top$ and $Y := \sum_{k=1}^{b-1} X_k$ thus the vectors z^k are sampled with replacement from $\{e_1, \ldots, e_{i-1}, e_{i+1}, \ldots, e_n\}$. Consequently

$$\mathbb{P}\left[z_j^k = 1\right] = \frac{1}{n-1} \ , \quad \forall j \in \{1, \ldots, i-1, i+1, \ldots, n\} \ .$$

We are now in a position to apply the Bernstein matrix inequality. To this end we have

- A sum of centered random matrices: $\mathbb{E}\left[X_k\right] = 0$.

- Let k^* be the unique index such that $z_{k^*}^k = 1$. We have a uniform bound of the largest eigenvalue of our X_k

$$\begin{aligned}
\lambda_{\max}(X_k) &= \frac{1}{b}\lambda_{\max}\left(\sum_{j\in[n]\setminus\{i\}} z_j^k a_j a_j^\top - \frac{1}{n-1}\sum_{j\in[n]\setminus\{i\}} a_j a_j^\top\right) \\
&\leq \frac{1}{b}\lambda_{\max}\left(\sum_{j\in[n]\setminus\{i\}} z_j^k a_j a_j^\top\right) \\
&= \frac{1}{b}\lambda_{\max}\left(a_{k^*} a_{k^*}^\top\right) \\
&\leq \frac{1}{b}\frac{L_{\max}}{U} \ ,
\end{aligned} \tag{35}$$

  where we applied the Lemma 9 in the first inequality.

- And a bound on the variance too

$$\begin{aligned}
\mathbb{E}\left[X_k^2\right] &= \mathbb{E}\left(\frac{1}{b}\sum_{j\in[n]\setminus\{i\}}\left(z_j^k - \frac{1}{n-1}\right) a_j a_j^\top\right)^2 \\
&= \frac{1}{b^2}\mathbb{E}\left(\sum_{j,p\in[n]\setminus\{i\}} z_j^k z_p^k a_j a_j^\top a_p a_p^\top - \frac{2}{n-1}\sum_{j,p\in[n]\setminus\{i\}} z_j^k a_j a_j^\top a_p a_p^\top + \frac{1}{(n-1)^2}\sum_{j,p\in[n]\setminus\{i\}} a_j a_j^\top a_p a_p^\top\right) \\
&= \frac{1}{b^2}\sum_{j,p\in[n]\setminus\{i\}}\left(\left(\mathbb{E}\left[z_j^k z_p^k\right] a_j a_j^\top a_p a_p^\top - \frac{2}{n-1}\mathbb{E}\left[z_j^k\right] a_j a_j^\top a_p a_p^\top + \frac{1}{(n-1)^2} a_j a_j^\top a_p a_p^\top\right)\right. \\
&= \frac{1}{b^2}\sum_{j,p\in[n]\setminus\{i\}}\left(\mathbb{E}\left[z_j^k z_p^k\right] a_j a_j^\top a_p a_p^\top - \frac{2}{(n-1)^2} a_j a_j^\top a_p a_p^\top + \frac{1}{(n-1)^2} a_j a_j^\top a_p a_p^\top\right) \\
&= \frac{1}{b^2}\left(\frac{1}{n-1}\sum_{j\in[n]\setminus\{i\}} a_j a_j^\top a_j a_j^\top - \frac{1}{(n-1)^2}\sum_{j,p\in[n]\setminus\{i\}} a_j a_j^\top a_p a_p^\top\right) \ ,
\end{aligned} \tag{36}$$

where, in the last equality, we used that $z_j^k z_p^k = 0$ if $j \neq p$ and $\mathbb{E}\left[z_j^k z_j^k\right] = \mathbb{E}\left[z_j^k\right] = \frac{1}{n-1}$, so that

$$\begin{aligned}\sum_{j,p\in[n]\setminus\{i\}} \mathbb{E}\left[z_j^k z_p^k\right] a_j a_j^\top a_p a_p^\top &= \mathbb{E}\left[\sum_{j,p\in[n]\setminus\{i\}} z_j^k z_p^k\right] a_j a_j^\top a_p a_p^\top \\ &= \sum_{j\in[n]\setminus\{i\}} \mathbb{E}\left[z_j^k z_j^k\right] a_j a_j^\top a_p a_p^\top \\ &= \frac{1}{n-1}\sum_{j\in[n]\setminus\{i\}} a_j a_j^\top a_p a_p^\top \ .\end{aligned}$$

Summing in (36), taking the largest eigenvalue and applying Lemma 9 results in

$$\begin{aligned}\lambda_{\max}\left(\sum_{k=1}^{b-1}\mathbb{E}\left[X_k^2\right]\right) &\leq \lambda_{\max}\left(\sum_{k=1}^{b-1}\frac{1}{b^2}\frac{1}{n-1}\sum_{j\in[n]\setminus\{i\}} a_j a_j^\top a_j a_j^\top\right) \\ &\leq \frac{b-1}{b^2}\left(\max_{j\in[n]\setminus\{i\}}\lambda_{\max}\left(a_j a_j^\top\right)\right)\cdot\lambda_{\max}\left(\frac{1}{n-1}\sum_{j\in[n]\setminus\{i\}} a_j a_j^\top\right) \\ &\leq \frac{b-1}{b^2}\frac{L_{\max}}{U^2}L_{[n]\setminus\{i\}} \ . \end{aligned} \tag{37}$$

Considering Equations (35) and (37) and applying the matrix Bernstein concentration inequality in Theorem 7 we get

$$U\mathbb{E}\left[\lambda_{\max}(N)\right] \leq \sqrt{2\frac{b-1}{b^2}L_{\max}L_{[n]\setminus\{i\}}\log d} + \frac{1}{3}\frac{L_{\max}}{b}\log d$$

Taking the maximum over i and using $L_{[n]\setminus\{i\}} \leq \frac{n}{n-1}L$ we have that

$$\max_{i=1,\ldots,n} U\mathbb{E}\left[\lambda_{\max}(N)\right] \leq \sqrt{2\frac{b-1}{b^2}\frac{n}{n-1}L_{\max}L\log d} + \frac{1}{3}\frac{L_{\max}}{b}\log d$$

Combining the above result with (32) leads us to

$$\begin{aligned}\mathcal{L} &\leq \frac{(n-b)L_{\max}}{b(n-1)} + \frac{n(b-1)L}{b(n-1)} + \sqrt{2\left(\frac{b-1}{b}\frac{n}{n-1}L\right)\cdot\left(\frac{1}{b}L_{\max}\log d\right)} + \frac{1}{3}\frac{L_{\max}}{b}\log d \\ &\leq \frac{(n-b)L_{\max}}{b(n-1)} + \frac{n(b-1)L}{b(n-1)} + \frac{b-1}{b}\frac{n}{n-1}L + \frac{4}{3}\frac{L_{\max}}{b}\log(d) \\ &= 2\frac{b-1}{b}\frac{n}{n-1}L + \frac{1}{b}\left(\frac{4}{3}\log(d) + \frac{n-b}{n-1}\right)L_{\max}.\end{aligned}$$

where in the second inequality we used the inequality $\sqrt{2ab} \leq a + b$. □

B. Linear Algebra Tools

This appendix is dedicated to the presentation of useful results to manipulate more easily the smoothness constants.

B.1. Spectral Lemmas

Let us recall some useful spectral results on Hermitian and positive semi-definite matrices.

Lemma 6. *(Weyl's inequality) Let $A, B \in \mathbb{R}^{n\times n}$ symmetric matrices. Assume that the eigenvalues of A (resp. B) are sorted* i.e., *$\lambda_1(A) \geq \cdots \geq \lambda_n(A)$ (resp. $\lambda_1(B) \geq \cdots \geq \lambda_n(B)$). Then, we have*

$$\lambda_{i+j-1}(A+B) \leq \lambda_i(A) + \lambda_j(B) \ . \tag{38}$$

whenever $i, j \geq 1$ and $i + j - 1 \leq n$.

Moreover, as a direct consequence of the variational characterization of eigenvalues, namely

$$\lambda_{\max}(A) = \max_{v \neq 0} \frac{v^\top A v}{\|v\|_2^2} \ , \tag{39}$$

we have an inequality between the maximum diagonal term of a positive semi-definite matrices and its maximum eigenvalue.

Lemma 7. *Let $A \in \mathbb{R}^{n \times n}$ positive semi-definite matrix and the vector containing its diagonal $d := \mathrm{diag}(A)$. Then, we have*

$$\max_{i=1,\dots,n} d_i \leq \lambda_{\max}(A) \ . \tag{40}$$

The following lemma is a direct consequence of Weyl's inequality for $i = j = 1$.

Lemma 8. *Let $A, B \in \mathbb{R}^{n \times n}$ symmetric matrices. Then, we have*

$$\lambda_{\max}(A + B) \leq \lambda_{\max}(A) + \lambda_{\max}(B) \ . \tag{41}$$

Lastly, we present a result arising from previous lemma.

Lemma 9. *Let $A, B \in \mathbb{R}^{n \times n}$ symmetric matrices such that B is positive semi-definite. Then, we have*

$$\lambda_{\max}(A - B) \leq \lambda_{\max}(A) \ . \tag{42}$$

Proof. Let $A, B \in \mathbb{R}^{n \times n}$ symmetric matrices such that B is positive semi-definite. We get directly

$$\begin{aligned} \lambda_{\max}(A - B) &\leq \lambda_{\max}(A) + \lambda_{\max}(-B) \\ &= \lambda_{\max}(A) - \lambda_{\min}(B) \\ &\leq \lambda_{\max}(A) \ , \end{aligned}$$

where the first inequality stems from Lemma 8 and the second from $B \succeq 0$. □

B.2. Basic properties of the smoothness constants

The complexity results of Gower et al. (2018) depends on smoothness constants defined in Section 3.1. Here are some inequalities giving an idea of the order of those constants.

Lemma 10. *Let $\emptyset \neq B \subseteq [n] = \{1, \dots, n\}$ a batch set drawn randomly without replacement. The following inequalities hold*

(i)

$$L_i \leq L_{\max} \quad \forall i = 1, \dots, n \ . \tag{43}$$

(ii)

$$L_B \leq \frac{1}{|B|} \sum_{i \in B} L_i \quad \forall i = 1, \dots, n \ . \tag{44}$$

(iii)

$$L \overset{(a)}{\leq} \bar{L} \overset{(b)}{\leq} L_{\max} \overset{(c)}{\leq} nL \overset{(d)}{\leq} n\bar{L} \ . \tag{45}$$

Proof. (i) One directly gets that $L_i \leq \max_{j=1,\dots,n} L_j = L_{\max}$.

(ii) This inequality states that the smoothness constant L_B of the averaged function f_B is upper bounded by the average of the corresponding smoothness constants L_i, over the batch B. The proof consists in $|B|$ repetitive calls of Lemma 8.

(iii) (a) Direct implication of (ii) for $B = [n]$.

(b) Direct calculation

$$\bar{L} = \frac{1}{n}\sum_{i=1}^{n} L_i \leq \frac{1}{n}\sum_{i=1}^{n} L_{\max} \leq L_{\max} \ .$$

(c) Let us first recall the matrix formulation of our smoothness constants:

$$L = \frac{U}{n}\lambda_{\max}(AA^\top) = \frac{U}{n}\lambda_{\max}(A^\top A)$$

and

$$L_{\max} = U \max_{i=1,\dots,n} e_i^\top A^\top A e_i \ ,$$

Using the min-max theorem, we have that

$$\lambda_{\max}(A^\top A) = \max_{x\neq 0} \frac{x^\top A^\top A x}{\|x\|_2^2} \geq \max_{i=1,\dots,n} e_i^\top A^\top A e_i.$$

Dividing the above by n on both sides gives

$$L \geq \frac{L_{\max}}{n} \ .$$

(d) Direct consequence of (a).

□

C. Matrix Bernstein Inequality: Sampling Without Replacement

In this appendix, we present the matrix Bernstein inequality for independent Hermitian matrices from Tropp (2015). We also provide another version of this theorem for matrices sampled *without* replacement and prove it as explicitly as possible, taking our inspiration from Tropp (2011). The proof is based the possibility of transferring the results from sampling *with* to *without* through the inequality (50) due to Gross & Nesme (2010). The exact same work can be done for the tail bound, which is for instance used in Bach (2013).

C.1. Original Bernstein inequality for independent matrices

We first present Theorem 4 which gives a Bernstein inequality for a sum of random and independent Hermitian matrices whose eigenvalues are upper bounded. If the matrices X_k are sampled from a finite set $\mathcal{X}$, one can interpret this random sampling of independent matrices as a random sampling *with* replacement.

Theorem 4 (Tropp (2015), Theorem 6.6.1: Matrix Bernstein Inequality)**.** *Consider a finite sequence $\{X_k\}_{k=1,\dots,n}$ of n independent, random, Hermitian matrices with dimension d. Assume that*

$$\mathbb{E}\, X_k = 0 \quad \text{and} \quad \lambda_{\max}(X_k) \leq L \quad \textit{for each index } k \ .$$

Introduce the random matrix

$$S_X := \sum_{k=1}^{n} X_k \ .$$

Let $v(S_X)$ be the matrix variance statistic of the sum:

$$v(S_X) := \left\|\mathbb{E}\, S_X^2\right\| = \left\|\sum_{k=1}^{n} \mathbb{E}\, X_k^2\right\| = \lambda_{\max}\left(\sum_{k=1}^{n} \mathbb{E}\, X_k^2\right) \ . \tag{46}$$

Then

$$\mathbb{E}\, \lambda_{\max}(S_X) \leq \sqrt{2v(S_X)\log d} + \frac{1}{3}L\log d \ . \tag{47}$$

This theorem is the one we extend in Theorem 7 to the case when the random matrices X_k are sampled *without* replacement from a finite set $\mathcal{X}$. We drew our inspiration from the proof of the matrix Chernoff inequality in Tropp (2011) and the one of the matrix Bernstein tail bound in Bach (2013), both in the case of sampling without replacement.

C.2. Technical random matrices prerequisites

Before proving Theorem 7, which extends the matrix Bernstein inequality to sampling without replacement, we need to introduce the key tools of the matrix Laplace transform technique. This technique is precious to prove tail bounds for sums of random matrices such as Chernoff, Hoeffding or Bernstein bounds, as presented in (Tropp, 2012).

Here, $\|\cdot\|$ denotes the spectral norm, which is defined for any Hermitian matrix H by

$$||M|| = \max\left\{\lambda_{\max}(H), -\lambda_{\min}(H)\right\} \quad . \tag{48}$$

We also introduce the moment generating function (mgf) and the cumulant generating function (cgf) of a random matrix, which are essential in the Laplace transform method approach.

Definition 6 (Matrix Mgf and Cgf)**.** Let X be a random Hermitian matrix. For all $\theta \in \mathbb{R}$, the matrix generating function M_X and the matrix cumulant generating function Ξ_X are given by

$$M_X(\theta) := \mathbb{E}\, e^{\theta X}$$

and

$$\Xi_X(\theta) := \log \mathbb{E}\, e^{\theta X} \quad .$$

Remark 4. These expectations may not exist for all values of θ.

Proposition 2 (Tropp (2015), Proposition 3.2.2: Expectation Bound of the Maximum Eigenvalue)**.** *Let X be a random Hermitian matrix. Then*

$$\mathbb{E}\,\lambda_{\max}(X) \leq \inf_{\theta>0}\left\{\frac{1}{\theta}\log \mathbb{E}\,\mathrm{tr}\,\mathrm{e}^{\theta X}\right\} \quad . \tag{49}$$

Remark 5. This proposition is an adaptation of the Laplace transform method to obtain a bound of the expectation of the maximum eigenvalue of a random Hermitian matrix. Contrary to the tail bounds, there is no exact analog of the expectation bounds in the scalar setting.

Proof of Proposition 2. Fix a positive number θ. Because $\lambda_{\max}(\cdot)$ is a positive-homogeneous map, we have

$$\begin{aligned}
\mathbb{E}\,\lambda_{\max}(X) &= \frac{1}{\theta}\mathbb{E}\,\lambda_{\max}(\theta X) \\
&= \frac{1}{\theta}\mathbb{E}\,\log \mathrm{e}^{\lambda_{\max}(\theta X)} \\
&\leq \frac{1}{\theta}\log \mathbb{E}\,\mathrm{e}^{\lambda_{\max}(\theta X)} \\
&= \frac{1}{\theta}\log \mathbb{E}\,\lambda_{\max}\left(\mathrm{e}^{\theta X}\right) \\
&\leq \frac{1}{\theta}\log \mathbb{E}\,\mathrm{tr}\,\mathrm{e}^{\theta X} \quad ,
\end{aligned}$$

where in the third line we used the Jensen's inequality, in the fourth one the spectral mapping theorem and in the last line the domination by the trace of a positive-definite matrix. □

Theorem 5 (Tropp (2015), Theorem 8.1.1: Lieb)**.** *Let H be a fixed Hermitian matrix with dimension d. The function*

$$X \to \mathrm{tr}\exp\left(H + X\right)$$

is a concave map on the the convex cone of $d \times d$ positive-definite matrices.

Proof of Theorem 5. See Chapter 8 in Tropp (2015). □

Corollary 1. *Let H be a fixed Hermitian matrix with dimension d. Let X be a random Hermitian matrix of same dimension. The following inequality holds*

$$\mathbb{E}\,\mathrm{tr}\exp\left(H + X\right) \leq \mathrm{tr}\exp\left(H + \log \mathbb{E}\, e^{X}\right)$$

is a concave map on the the convex cone of $d \times d$ positive-definite matrices.

Proof of Corollary 1. Introducing $Y = e^X$, we have directly

$$\begin{aligned}\mathbb{E}\,\mathrm{tr}\exp\left(H+X\right) &= \mathbb{E}\,\mathrm{tr}\exp\left(H+\log e^X\right)\\ &= \mathbb{E}\,\mathrm{tr}\exp\left(H+\log Y\right)\\ &\leq \mathrm{tr}\exp\left(H+\log\mathbb{E}\,Y\right)\\ &= \mathrm{tr}\exp\left(H+\log\mathbb{E}\,e^X\right)\quad .\end{aligned}$$

where the inequality comes from the application of Theorem 5 and Jensen's inequality. □

Lemma 11 (Tropp (2015), Lemma 3.5.1 or Tropp (2012), Lemma 3.4: Subadditivity of Matrix Cgfs)**.** *Consider a finite sequence $\{X_k\}$ of independent, random, Hermitian matrices of the same dimension. Let $\theta \in \mathbb{R}$, then*

$$\begin{aligned}\mathrm{tr}\exp\left(\Xi_{\sum k=1^n X_k}(\theta)\right) &= \mathbb{E}\,\mathrm{tr}\exp\left(\theta\sum\nolimits_k X_k\right)\\ &\leq \mathrm{tr}\exp\left(\sum\nolimits_k \log\mathbb{E}\,\mathrm{e}^{\theta X_k}\right)\\ &= \mathrm{tr}\exp\left(\sum\nolimits_k \Xi_{X_k}(\theta)\right)\quad .\end{aligned}$$

Proof of Lemma 11. Let us assume, without loss of generality, that $\theta = 1$. Let a finite sequence $\{X_k\}_{k=1}^n$ of n independent, random, Hermitian matrices of the same dimension. We write down $\mathbb{E}_k$ the expectation with respect only to the k-th random matrix X_k.

$$\begin{aligned}\mathrm{tr}\exp\left(\Xi_{\sum_{k=1}^n X_k}(1)\right) &= \mathrm{tr}\exp\left(\log\mathbb{E}\,\exp\left(\sum_{k=1}^n X_k\right)\right)\\ &= \mathbb{E}\,\mathrm{tr}\exp\left(\sum_{k=1}^n X_k\right)\\ &= \mathbb{E}_1\ldots\mathbb{E}_{n-1}\mathbb{E}_n\,\mathrm{tr}\exp\left(\sum_{k=1}^{n-1} X_k + X_{n+1}\right)\\ &\leq \mathbb{E}_1\ldots\mathbb{E}_{n-1}\,\mathrm{tr}\exp\left(\sum_{k=1}^{n-1} X_k + \log\mathbb{E}_n\,e^{X_{n+1}}\right)\\ &= \mathbb{E}_1\ldots\mathbb{E}_{n-1}\,\mathrm{tr}\exp\left(\sum_{k=1}^{n-2} X_k + X_{n-1} + \log\mathbb{E}_n\,e^{X_{n+1}}\right)\\ &\leq \mathbb{E}_1\ldots\mathbb{E}_{n-2}\,\mathrm{tr}\exp\left(\sum_{k=1}^{n-2} X_k + \log\mathbb{E}_{n-1}\,e^{X_{n-1}} + \log\mathbb{E}_n\,e^{X_n}\right)\\ &\leq \cdots \leq \mathrm{tr}\exp\left(\sum\nolimits_k \log\mathbb{E}\,\mathrm{e}^{\theta X_k}\right)\\ &= \mathrm{tr}\exp\left(\sum_k \Xi_{X_k}(\theta)\right)\quad .\end{aligned}$$

where first and second inequalities result from Corollary 1, the last one comes the fact that $\mathbb{E}_k\,e^{X_k} = \mathbb{E}\,e^{X_k}, \forall k \in [n]$ and the final equality directly comes from an indentification of Definition 6. □

Lemma 12 (Tropp (2015), Lemma 6.6.2: Matrix Bernstein Mgf and Cgf Bounds)**.** *Let X a random Hermitian matrix such that*

$$\mathbb{E}\,X = 0 \quad \textit{and} \quad \lambda_{\max}(X) \leq L\ .$$

Then, for $0 < \theta < 3/L$,

$$M_X(\theta) := \mathbb{E}\,e^{\theta X} \preceq \exp\left(\frac{\theta^2/2}{1-\theta L/3}\cdot\mathbb{E}\,X^2\right)$$

and

$$\Xi_X(\theta) := \log \mathbb{E}\, e^{\theta X} \preceq \frac{\theta^2/2}{1-\theta L/3} \cdot \mathbb{E}\, X^2 \quad .$$

Proof of Lemma 12. See Tropp (2015). □

C.3. Extended results for sampling without replacement

This section is dedicated to the main result, Lemma 13, needed for transferring results from sampling *with* to *without* replacement. This lemma is actually the matrix version of a classical result from Hoeffding (1963). We then combine it with previous results of Appendix C.2 to produce a new master bound in Theorem 6, which is the key inequality of the proof of Theorem 7.

Lemma 13 (Gross & Nesme (2010), Domination of the Trace of the Mgf of a Sample Without Replacement)**.** *Consider two finite sequences, of same length n, $\{X_k\}_{k=1,\dots,n}$ and $\{Y_k\}_{k=1,\dots,n}$ of Hermitian random matrices sampled respectively* with *and* without *replacement from a finite set $\mathcal{X}$. Let $\theta \in \mathbb{R}$, $S_X := \sum_{k=1}^n X_k$ and $S_Y := \sum_{k=1}^n Y_k$, then*

$$\operatorname{tr} M_{S_Y}(\theta) := \mathbb{E} \operatorname{tr} \exp(\theta S_Y) \leq \mathbb{E} \operatorname{tr} \exp(\theta S_X) \quad . \tag{50}$$

Proof of Lemma 13. The left-hand side equality directly arises from Definition 6 and the fact that the trace commutes with the expectation because it is a linear operator. For the right-hand side inequality, see the proof in Gross & Nesme (2010). □

Theorem 6 (Master Bound for a Sum of Random Matrices Sampled Without Replacement)**.** *Consider two finite sequences, of same length n, $\{X_k\}_{k=1,\dots,n}$ and $\{Y_k\}_{k=1,\dots,n}$ of Hermitian random matrices of same size sampled respectively* with *and* without *replacement from a finite set $\mathcal{X}$. Then*

$$\mathbb{E}\, \lambda_{\max}\left(\sum_{k=1}^n Y_k\right) \leq \inf_{\theta>0} \left\{ \frac{1}{\theta} \log \operatorname{tr} \exp\left(\sum_{k=1}^n \log \mathbb{E}\, \mathrm{e}^{\theta X_k}\right)\right\} \quad . \tag{51}$$

Remark 6. This theorem is a modified version of Theorem 3.6.1 in Tropp (2015) for a sum of matrices sampled without replacement.

Proof of Theorem 6. Consider two finite sequences, of same length, $\{X_k\}$ and $\{Y_k\}$ of Hermitian random matrices of same size sampled respectively *with* and *without* replacement from a finite set $\mathcal{X}$. Let θ a positive number.

$$\begin{aligned}\mathbb{E}\, \lambda_{\max}\left(\sum_{k=1}^n Y_k\right) &\leq \inf_{\theta>0} \left\{ \frac{1}{\theta} \log \mathbb{E} \operatorname{tr} \exp\left(\theta \sum_{k=1}^n Y_k\right)\right\} \leq \inf_{\theta>0} \left\{ \frac{1}{\theta} \log \mathbb{E} \operatorname{tr} \exp\left(\theta \sum_{k=1}^n X_k\right)\right\} \\ &\leq \inf_{\theta>0} \left\{ \frac{1}{\theta} \log \operatorname{tr} \exp\left(\sum_{k=1}^n \log \mathbb{E}\, \mathrm{e}^{\theta X_k}\right)\right\} \quad .\end{aligned}$$

where we used successively Proposition 2, Lemma 13 and Lemma 11. First, we use the expectation bound for the maximum eigenvalue. We then use the main result of Gross & Nesme (2010) and invoked in Tropp (2011) to extend the matrix Chernoff bound for matrices sampled *without* replacement. This lemma allows us to transfer our results to sampling *with* replacement. And finally, we then apply the subadditivity of matrix cgfs to get the desired result. □

C.4. Bernstein inequality for sampling without replacement

The following theorem is almost the same than Theorem 4, but in the case of matrices sampled *without* replacement from a finite set. The proof stems from results established in previous Appendices C.2 and C.3.

Theorem 7 (Matrix Bernstein Inequality Without Replacement)**.** *Let $\mathcal{X}$ be a finite set of Hermitian matrices with dimension d such that*

$$\lambda_{\max}(X) \leq L, \quad \forall X \in \mathcal{X} \quad .$$

Sample two finite sequences, of same length n, $\{X_k\}_{k=1,\dots,n}$ and $\{Y_k\}_{k=1,\dots,n}$ uniformly at random from $\mathcal{X}$ respectively with *and* without *replacement such that*

$$\mathbb{E}\, X_k = 0 \quad \forall k \quad .$$

Introduce the random matrices

$$S_X := \sum_{k=1}^{n} X_k \quad \textit{and} \quad S_Y := \sum_{k=1}^{n} Y_k \ .$$

Let $v(S_X)$ be the matrix variance statistic of the second sum

$$v(S_X) := \left\|\mathbb{E}\, S_X^2\right\| = \left\|\sum_{k=1}^{n} \mathbb{E}\, X_k^2\right\| = \lambda_{\max}\left(\sum_{k=1}^{n} \mathbb{E}\, X_k^2\right). \tag{52}$$

Then

$$\mathbb{E}\, \lambda_{\max}(S_Y) \leq \sqrt{2v(S_X)\log d} + \frac{1}{3} L \log d \ . \tag{53}$$

Proof of Theorem 7. Consider $\mathcal{X}$ a finite set of Hermitian matrices of dimension d such that

$$\lambda_{\max}(X) \leq L \quad \forall X \in \mathcal{X} \ .$$

Sample two finite sequences, of same length, $\{X_k\}$ and $\{Y_k\}$ uniformly at random from $\mathcal{X}$ respectively *with* and *without* replacement such that

$$\mathbb{E}\, X_k = 0 \quad \forall k \ .$$

The $\{X_k\}$ matrices are thus independent. Introduce the sums $S_X = \sum_{k=1}^n X_k$ and $S_Y = \sum_{k=1}^n Y_k$. Let us bound the expectation of the largest eigenvalue of the latter

$$\begin{aligned}
\mathbb{E}\, \lambda_{\max}(S_Y) = \mathbb{E}\, \lambda_{\max}\left(\sum_{k=1}^{n} Y_k\right) &\leq \inf_{\theta>0}\left\{\frac{1}{\theta}\log \operatorname{tr}\exp\left(\sum_{k=1}^{n}\log \mathbb{E}\, \mathrm{e}^{\theta X_k}\right)\right\} \\
&\leq \inf_{0<\theta<3/L}\left\{\frac{1}{\theta}\log \operatorname{tr}\exp\left(\frac{\theta^2/2}{1-\theta L/3}\sum_{k=1}^{n}\mathbb{E}\, X_k^2\right)\right\} \\
&\leq \inf_{0<\theta<3/L}\left\{\frac{1}{\theta}\log\left[d\,\lambda_{\max}\left(\exp\left(\frac{\theta^2/2}{1-\theta L/3}\mathbb{E}\, S_X^2\right)\right)\right]\right\} \\
&\leq \inf_{0<\theta<3/L}\left\{\frac{1}{\theta}\log\left[d\,\exp\left(\frac{\theta^2/2}{1-\theta L/3}\lambda_{\max}\left(\mathbb{E}\, S_X^2\right)\right)\right]\right\} \\
&\leq \inf_{0<\theta<3/L}\left\{\frac{1}{\theta}\log\left[d\,\exp\left(\frac{\theta^2/2}{1-\theta L/3}v(S_X)\right)\right]\right\} \\
&= \inf_{0<\theta<3/L}\left\{\frac{\log d}{\theta} + \frac{\theta/2}{1-\theta L/3}v(S_X)\right\} \ .
\end{aligned}$$

where the inequalities sucessively derive from Theorem 6, Lemma 12 combined with the monotony of $\operatorname{tr}\exp(\cdot)$, the fact that $\operatorname{tr}(M) \leq d\,\lambda_{\max}(M)$, $\forall M \in \mathbb{R}^{d\times d}$, the spectral mapping theorem and lastly (48) with $\mathbb{E}\, Y^2 \succeq 0$. Finally, one can complete the infimum, for instance using a computer algebra system, to finish the proof as it was stated in the original proof by Tropp (2015) [9]. In conclusion,

$$\mathbb{E}\, \lambda_{\max}(S_Y) \leq \sqrt{2v(S_X)\log d} + \frac{1}{3} L \log d \ .$$

□

D. Miscellaneous

Lemma 14 (Double counting). *Let $a_{i,C} \in \mathbb{R}$ for $i = 1, \ldots, n$ and $C \in \mathcal{C}$, where $\mathcal{C}$ is a collection of subsets of $[n]$. Then*

$$\sum_{C\in\mathcal{C}}\sum_{i\in C} a_{i,C} \quad = \quad \sum_{i=1}^{n}\sum_{C\in\mathcal{C}\,:\,i\in C} a_{i,C}. \tag{54}$$

[9] For instance : `Minimize[(log(d)/x) + ((x/2)/(1-(L/3)*x))*v, x >0, x < (3/L), x]` in Wolfram Alpha.

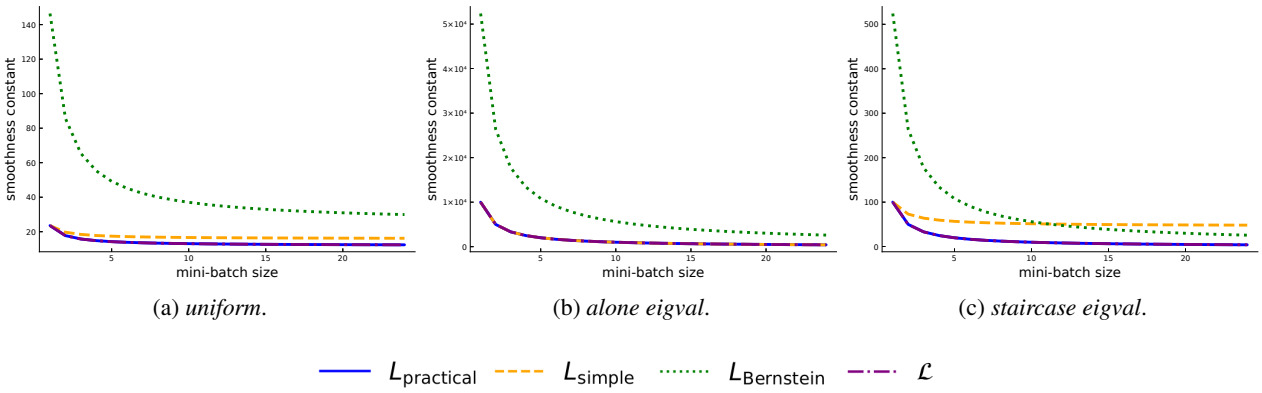
Algorithm 2 JACSKETCH PRACTICAL IMPLEMENTATION OF b-NICE SAGA

Input: mini-batch size b, step size $\gamma > 0$
Initialize: $w^0 \in \mathbb{R}^d$, $J^0 \in \mathbb{R}^{d\times n}$, $u^0 = \frac{1}{n} J^0 e$
for $k = 0, 1, 2, \ldots$ **do**
  Sample a fresh batch $B \subseteq [n]$ s.t. $|B| = b$
  $\text{aux} = \sum_{i\in B} (\nabla f_i(w^k) - J^k_{:i})$ // update the auxiliary vector
  $g^k = u^k + \frac{1}{b}\text{aux}$ // update the unbiased gradient estimate
  $u^{k+1} = u^k + \frac{1}{n}\text{aux}$ // update the biased gradient estimate
  $J^{k+1}_{:i} = \begin{cases} J^k_{:i} & i \notin B \\ \nabla f_i(w^k) & i \in B. \end{cases}$ // update the Jacobian estimate
  $w^{k+1} = w^k - \gamma g^k$ // take a step
end for

E. Additional Experiments

E.1. Experiment 1: estimates of the expected smoothness constant for artificial datasets

As described in Section 5, we compute our the *simple* and *Bernstein bounds*, our *practical estimate* and the true $\mathcal{L}$ for ridge regression applied to small artificial datasets: *uniform* ($n = 24, d = 50$), *staircase eigval* ($n = d = 24$) and *alone eigval* ($n = d = 24$). Figure 7 shows first that the *practical estimate* is a very close approximation of $\mathcal{L}$. On the one hand, we observe in Figure 7a that the *Bernstein bound* performs poorly since the feature dimension is very small $d = 50$. On the other hand, Figure 7c shows a regime change for $b \approx 10$, which highlight the usefulness of combining our bounds to approximate the expected smoothness constant. Finally, we observe that for the *alone eigval* dataset Figure 7b, which has one very large eigenvalue far from the rest of the spectrum, the *simple bound* matches $\mathcal{L}$ because the gap between $\bar{L}$ and L shrinks. Indeed, in this configuration $\bar{L} \approx L \approx \frac{L_{\max}}{n}$. When the spectrum is more concentrated, like for *staircase eigval*, we get a significant gap between $\bar{L}$ and L as shown in Figure 7c, where the *simple bound* is far from $\mathcal{L}$ when $b = n$.

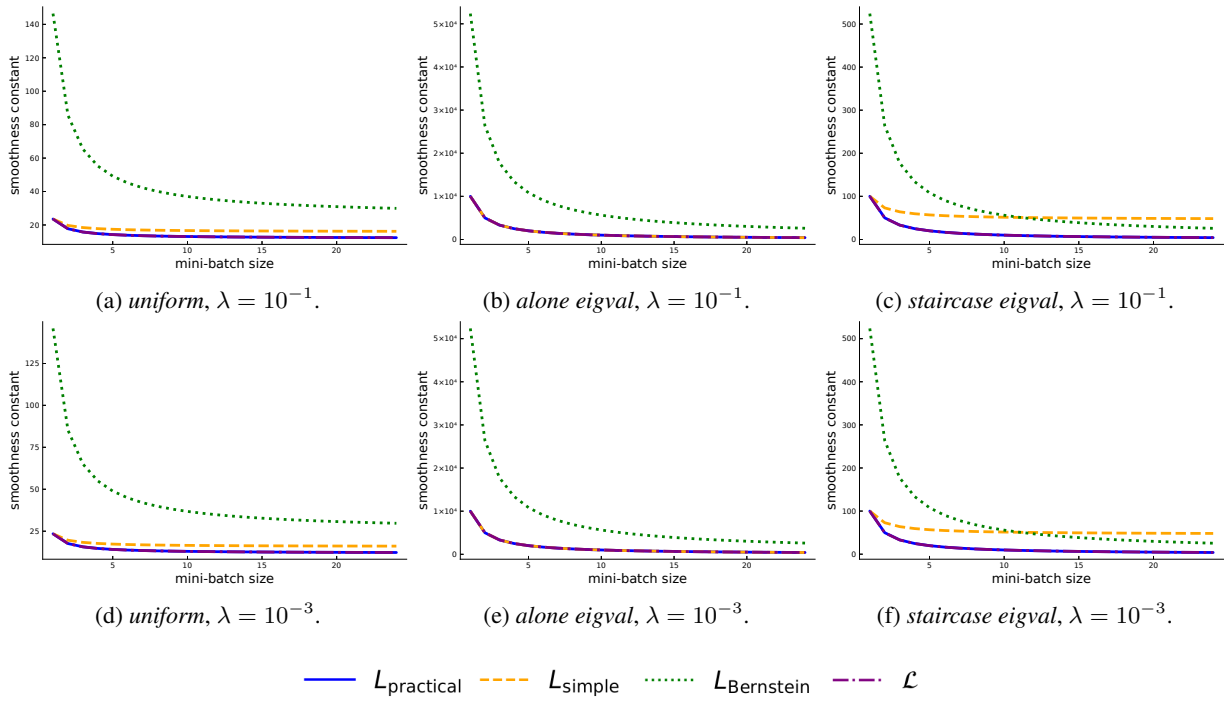


(a) *uniform*. (b) *alone eigval*. (c) *staircase eigval*.

Figure 7: Expected smoothness constant $\mathcal{L}$ and its upper-bounds the mini-batch size b varies (unscaled datasets, $\lambda = 10^{-1}$).

We also report the influence of changing the value of the regularization parameter λ. Figure 8 shows that this parameter has little impact on the general shape of the bounds and of $\mathcal{L}$.

Finally, we study the impact of scaling or standardizing (*i.e.,* removing the mean and dividing by the standard deviation for each feature) our artificial datasets. In order not to benefit from the diagonal shape of the *alone eigval* and *staircase eigval* datasets we also give examples of the bounds of $\mathcal{L}$ after a rotation of the data. The rotation aims at preserving the spectrum while erasing the diagonal structure of the covariance matrix $AA^\top$. This rotation procedure consists in transforming A into $Q^\top AQ$, where Q is the orthogonal matrix given by the QR decomposition of a random squared matrix (with dimension the same as the one of A) with uniformly random coefficients M, such that $M = QR$.



(a) *uniform*, $\lambda = 10^{-1}$. (b) *alone eigval*, $\lambda = 10^{-1}$. (c) *staircase eigval*, $\lambda = 10^{-1}$.

(d) *uniform*, $\lambda = 10^{-3}$. (e) *alone eigval*, $\lambda = 10^{-3}$. (f) *staircase eigval*, $\lambda = 10^{-3}$.

Figure 8: Expected smoothness constant $\mathcal{L}$ and its upper-bounds as a function of the mini-batch size for unscaled datasets with $\lambda = 10^{-1}$ (top) and $\lambda = 10^{-3}$ (bottom).

We observe in Figure 9 that rotations do not affect our estimates of $\mathcal{L}$, because they preserve the spectrum. Scaling non-diagonal datasets does not change the general shape neither. As predicted, scaling diagonal matrices leads to a particular case where the spectrum of the covariance matrix is flattened and for all $i \in [n]$, $L_i \approx L_{\max} \approx \bar{L}$. This is why we get a flat *simple bound* in Figures 9c and 9g. Even after those different types of preprocessing (rotation and scaling) and with different values of λ, we end up with the same strong observation that the *practical estimate* is a very sharp approximation of the expected smoothness constant.

E.2. Experiment 1: estimates of the expected smoothness constant for real datasets

In what folows, we also used publicly available datasets from LIBSVM[10] provided by Chang & Lin (2011) and from the UCI repository[11] provided by Dheeru & Karra Taniskidou (2017). We applied ridge regression to the following datasets: *YearPredictionMSD* ($n = 515,345, d = 90$) from LIBSVM and *slice* ($n = 53,500, d = 384$) from UCI. We also applied regularized logistic regression for binary classification on *ijcnn1* ($n = 141,691, d = 22$), *covtype.binary* ($n = 581,012, d = 54$), *real-sim* ($n = 72,309, d = 20,958$), *rcv1.binary* ($n = 697,641, d = 47,236$) and *news20.binary* ($n = 19,996, d = 1,355,191$) from LIBSVM. When a test set was available, we concatenated it with the train set to have more samples.

One can observe in Figure 10, that for unscaled datasets the *Bernstein bound* performs better than the *simple bound*, except for *YearPredictionMSD* ($n = 515,345, d = 90$) and *covtype.binary* ($n = 581,012, d = 54$). From Figure 11, we observe that after feature-scaling, the *Bernstein bound* is always below the *simple bound*.

E.3. Experiment 2: step size estimates for artificial datasets

In this section we give the step sizes estimate corresponding to the expected smoothness constant, the *simple* and *Bernstein* upper-bounds and the *practical* estimate for our small artificial datasets. In Figure 12, we show that the *practical* step size

[10] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

[11] https://archive.ics.uci.edu/ml/datasets/

estimate is larger than all others. Moreover, for except for small value sof b, our γ_{simple} or $\gamma_{\text{Bernstein}}$ estimates are larger than the one proposed in (Hofmann et al., 2015).

E.4. Experiment 2: step size estimates for real datasets

Here we show the step sizes estimate corresponding to the *simple* and *Bernstein* upper-bounds and the *practical* estimate for real datasets detailed in Appendix E.2. On these real data, unscaled in Figure 13 scaled in Figure 14, we see that the gap between our step size estimates and γ_{Hofmann} are even larger. We observe in Figure 13, accordlingly to previous remarks in Appendix E.2, that *simple bound* leads to higher step sizes than the *Bernstein* one. Yet, as noticed before, Figure 14 seems to show that scaling the data leads to $\gamma_{\text{Bernstein}}$ larger than γ_{simple}.

E.5. Experiment 3: comparison with previous SAGA settings

In this section we provide more example of the performance of our practical settings compared to previously known SAGA settings. In Figures 16 to 21 we run our experiments on real datasets introduced in detail in Appendix E.1. SAGA implementations are run until the suboptimality reaches a relative error of 10^{-4}, except in some cases where the Hofmann's exceeded our maximal number of epochs like in Figure 17. In Figure 19, the curves corresponding to Hofmann's settings are not displayed because it achieves a total complexity which is too large. Figure 15 shows an example of such a configuration.

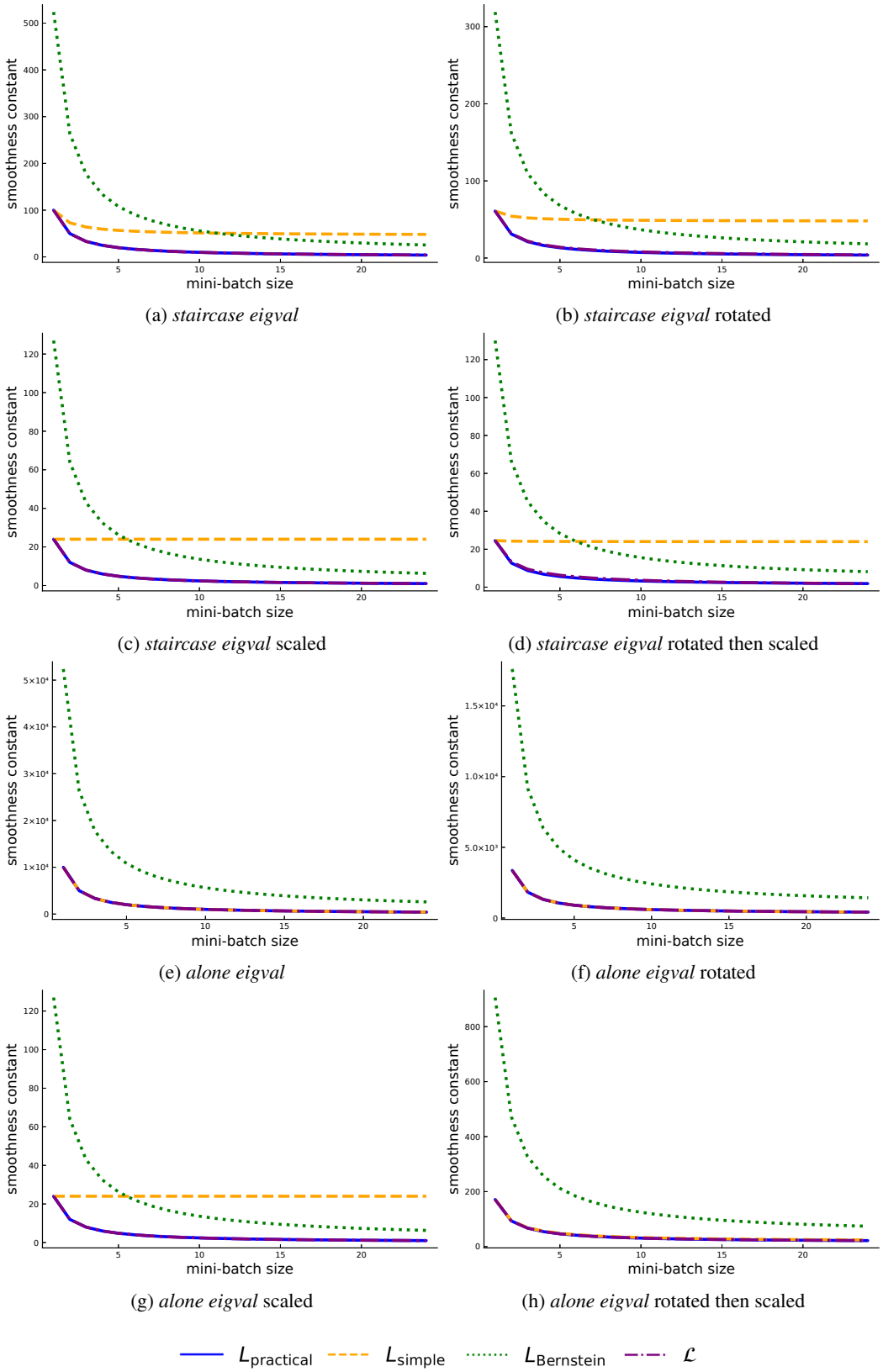
These experiments show that our settings $(b_{\text{practical}}, \gamma_{\text{practical}})$ outperforms whether the classical $(b = 1, \gamma_{\text{Defazio}})$ or the $(b = 20, \gamma_{\text{Hofmann}})$ settings both in terms of epochs and running time. Interestingly enough, Figure 21a exhibits a case for which our estimate of the optimal mini-batch size ($b_{\text{practical}} = 64, 141$) is closed to the number of data points ($n = 72,309$ for *real-sim*). So, our method also seems to indicate when to apply gradient descent rather than stochastic gradient methods.

E.6. Experiment 4: optimality of the mini-batch size

This experiment aims to estimate how close is our practical estimate $b_{\text{practical}}$ to the empirical best mini-batch size one could get running a grid search. We recall that we use the following grid for the mini-batch sizes: $\{2^i, i = 0, \dots, 14\}$, with $2^{16}, 2^{18}$ and n added in some cases. We show in the log-scaled Figures 22 to 27 the empirical total complexity K_{total}, *e.g.,* the number of computed gradients to reach a relative error of 10^{-4}, as a function of the mini-batch b.

We always observe a change of regime in the empirical complexity, except in Figure 27a. For small values of b, the complexity is of the same order of magnitude, then, for values greater than the empirical optimal mini-batch size, the complexity explodes. The exception of Figure 27a displays a case for which the total complexity is minimized for $b \approx n$, *e.g.,* for which classical gradient descent performs better than SAGA. Our *practical* settings succesfully predict this behaviour, like mentioned in Appendix E.5, since $b_{\text{practical}} = 64, 141 \approx n$.

This experiment shows that our optimal mini-batch size $b_{\text{practical}}$ correctly designates the largest mini-batch achieving the best complexity as large as possible, without reaching the regime where the total complexity explodes, or is predicting when to use gradient descent rather than stochastic methods.



(a) *staircase eigval*

(b) *staircase eigval* rotated

(c) *staircase eigval* scaled

(d) *staircase eigval* rotated then scaled

(e) *alone eigval*

(f) *alone eigval* rotated

(g) *alone eigval* scaled

(h) *alone eigval* rotated then scaled

Figure 9: Upper-bounds of the expected smoothness constant $\mathcal{L}$ for non-rotated (left) and rotated (right) datasets ($\lambda = 10^{-3}$).

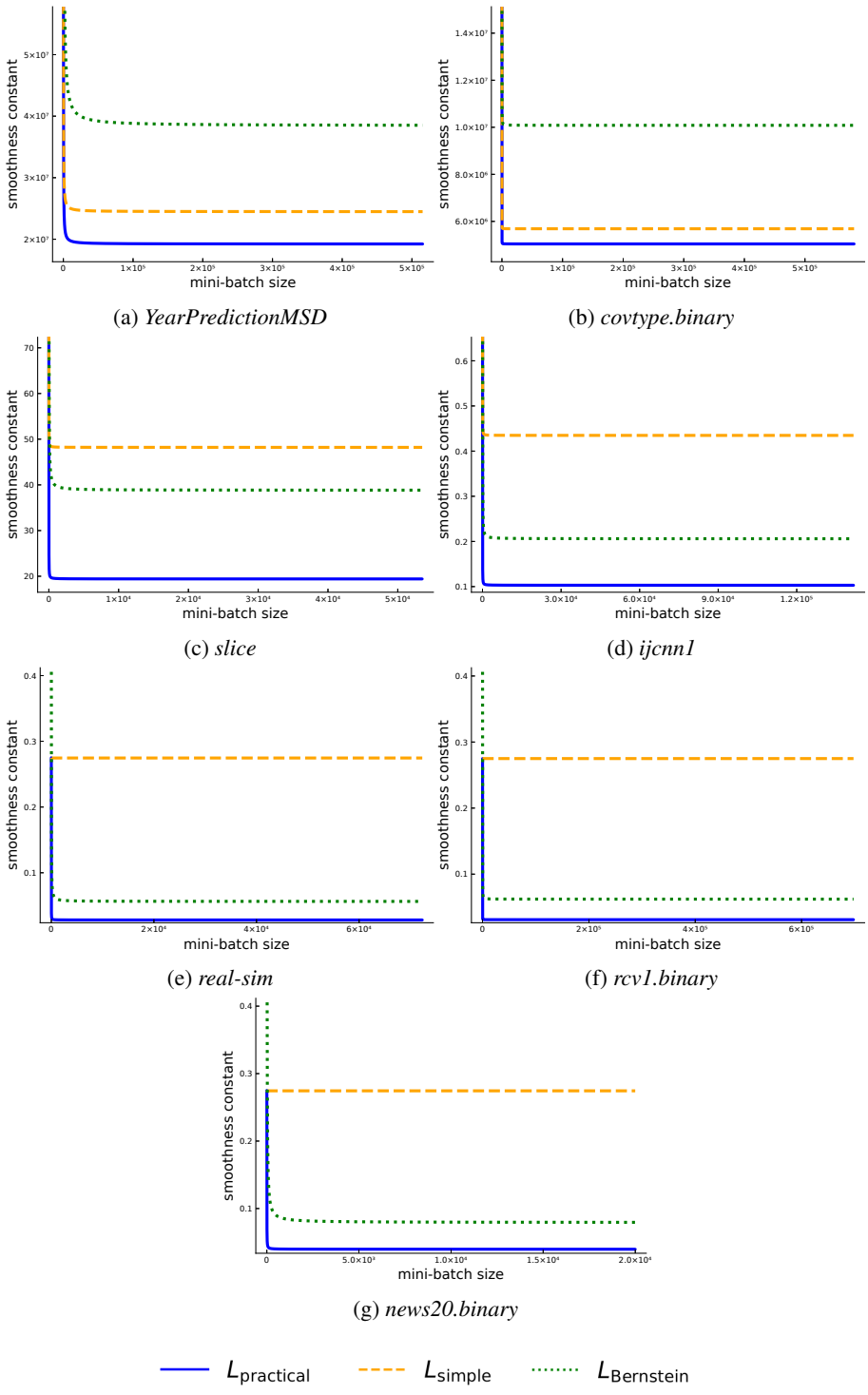
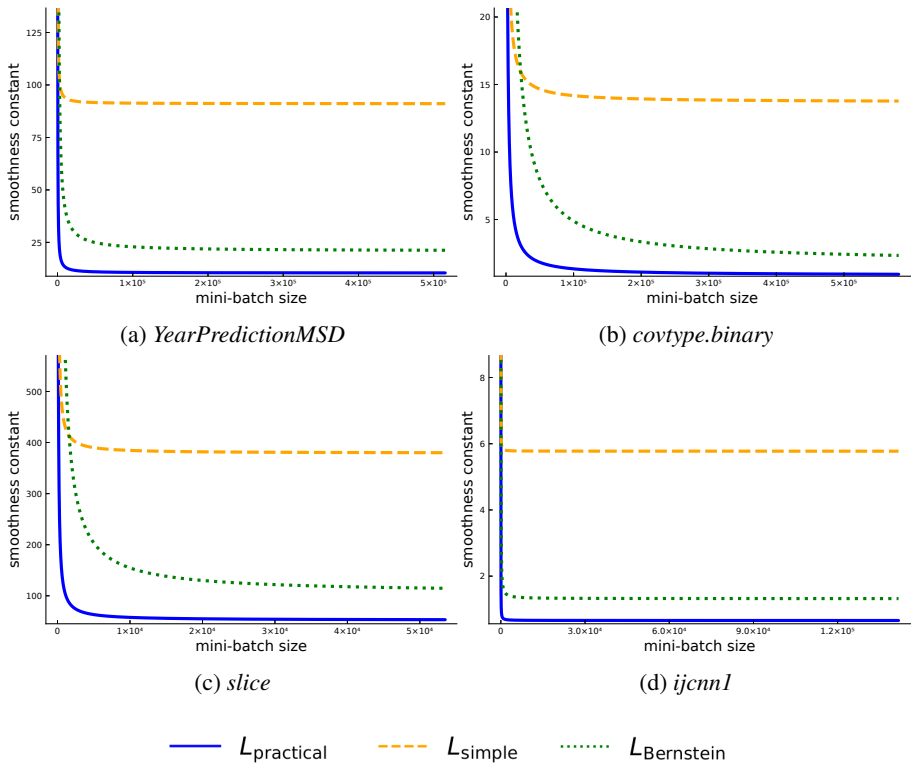


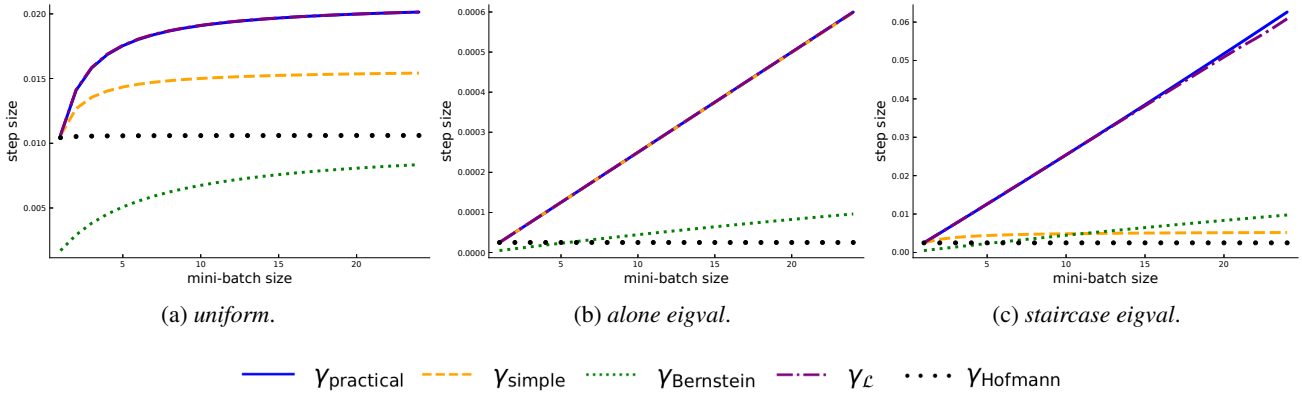
Figure 10: Upper-bounds of the expected smoothness constant for real unscaled datasets ($\lambda = 10^{-1}$).



(a) *YearPredictionMSD* (b) *covtype.binary*

(c) *slice* (d) *ijcnn1*

Figure 11: Upper-bounds of the expected smoothness constant of $\mathcal{L}$ for real feature-scaled datasets ($\lambda = 10^{-1}$).



(a) *uniform*. (b) *alone eigval*. (c) *staircase eigval*.

Figure 12: Step size estimates as a function the mini-batch size for unscaled artificial datasets ($\lambda = 10^{-1}$).

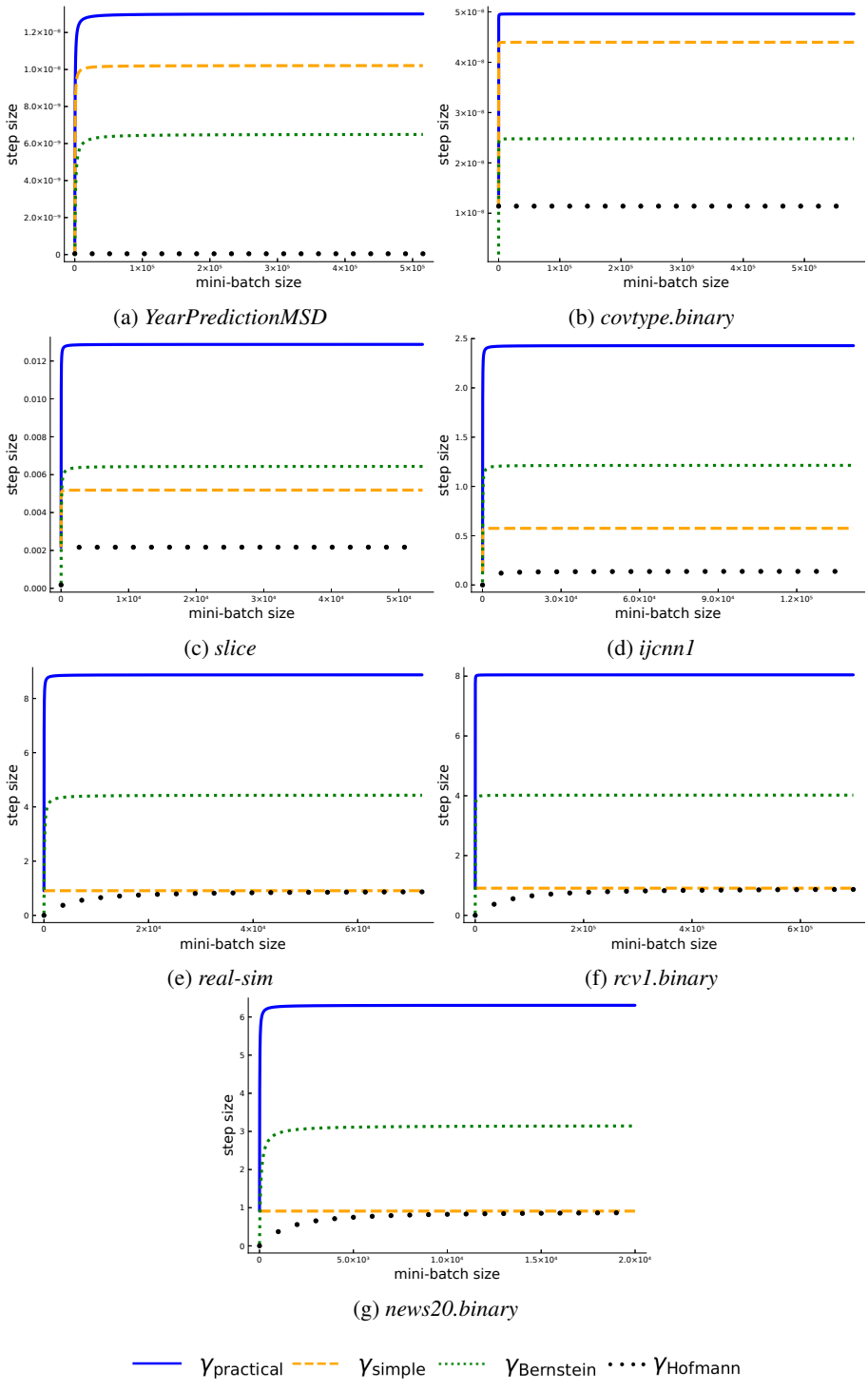


Figure 13: Step size estimates as a function the mini-batch size for real unscaled datasets ($\lambda = 10^{-1}$).

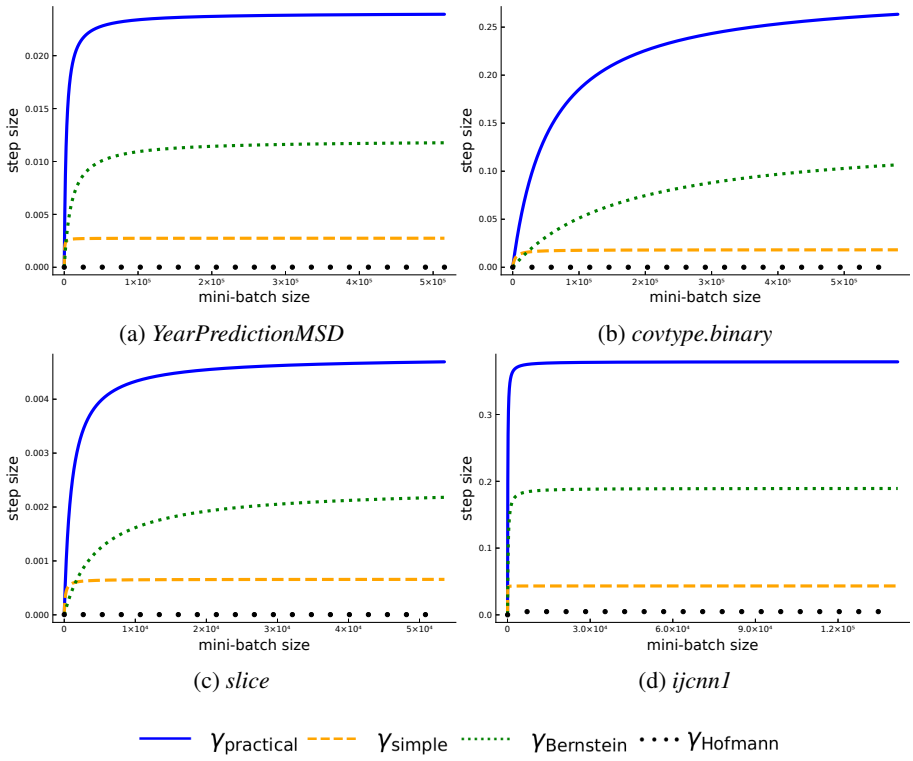


Figure 14: Step size estimates as a function the mini-batch size for real feature-scaled datasets ($\lambda = 10^{-1}$).

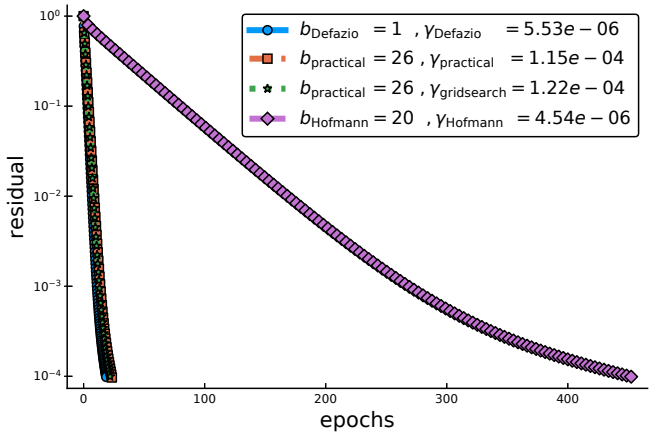
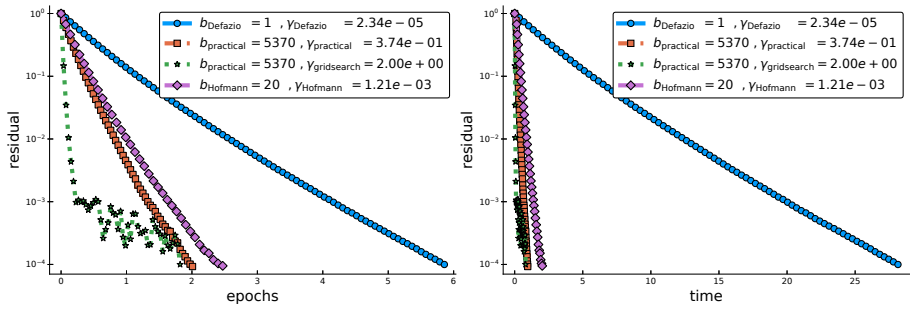
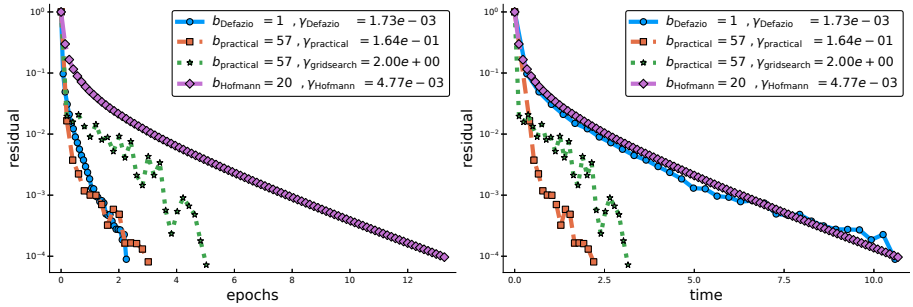


Figure 15: Poor performance of Hofmann's settings for the feature-scaled dataset *slice* ($\lambda = 10^{-1}$).

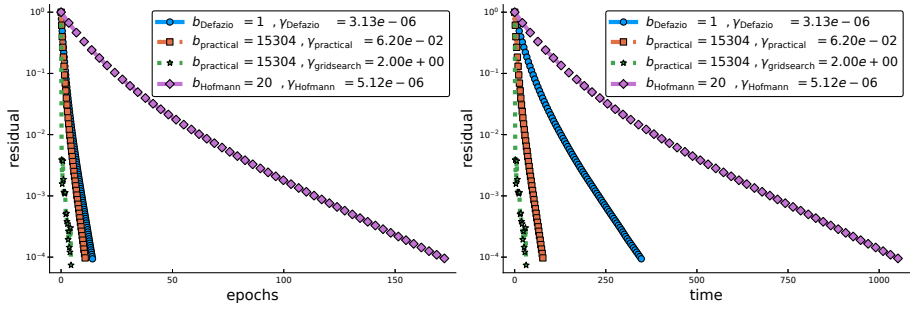


(a) $\lambda = 10^{-1}$

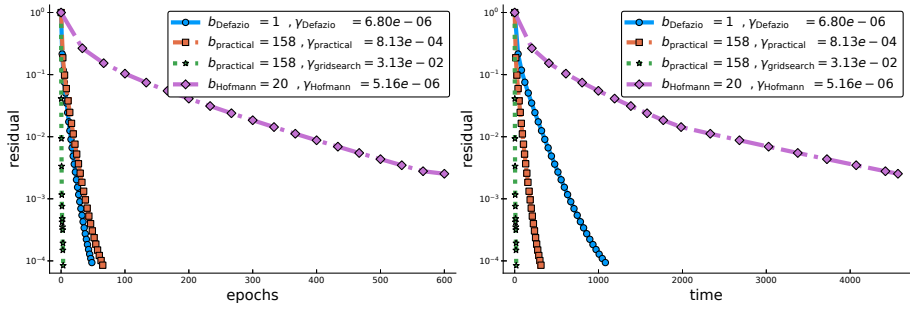


(b) $\lambda = 10^{-3}$

Figure 16: Performance of SAGA implementations for the feature-scaled dataset *ijcnn1*.

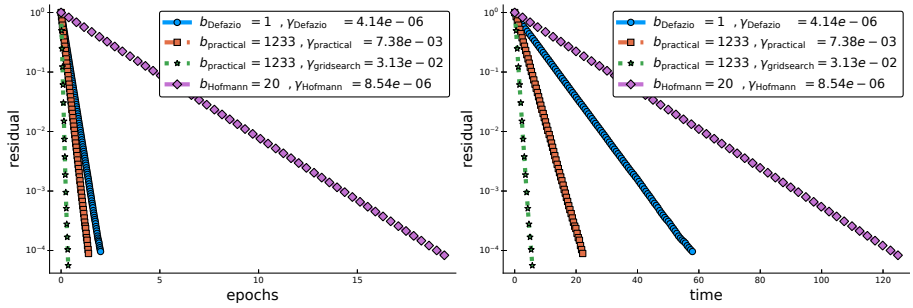


(a) $\lambda = 10^{-1}$

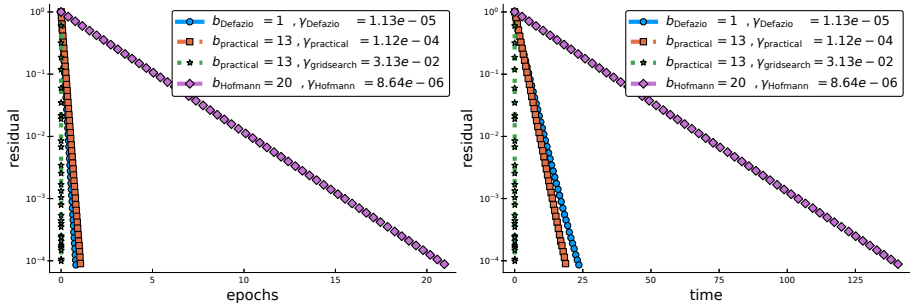


(b) $\lambda = 10^{-3}$

Figure 17: Performance of SAGA implementations for the feature-scaled dataset *covtype.binary*.

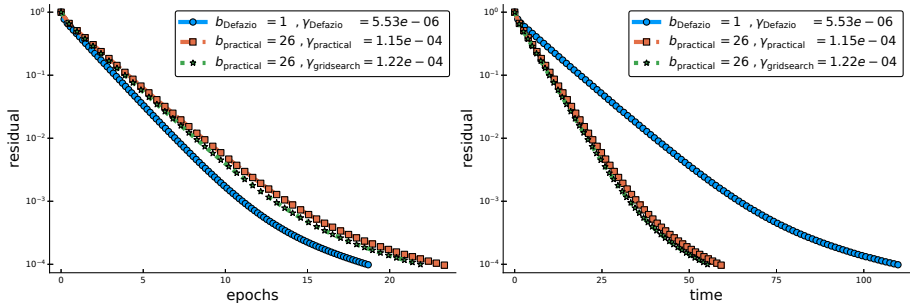


(a) $\lambda = 10^{-1}$

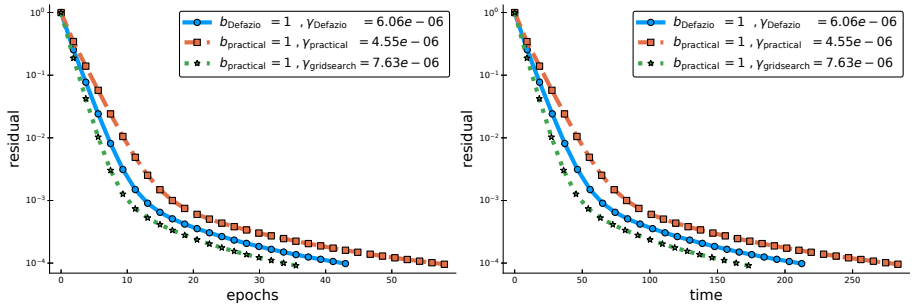


(b) $\lambda = 10^{-3}$

Figure 18: Performance of SAGA implementations for the feature-scaled dataset *YearPredictionMSD*.

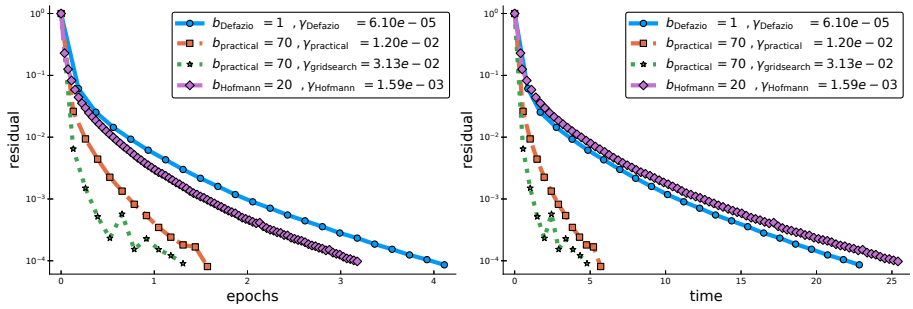


(a) $\lambda = 10^{-1}$

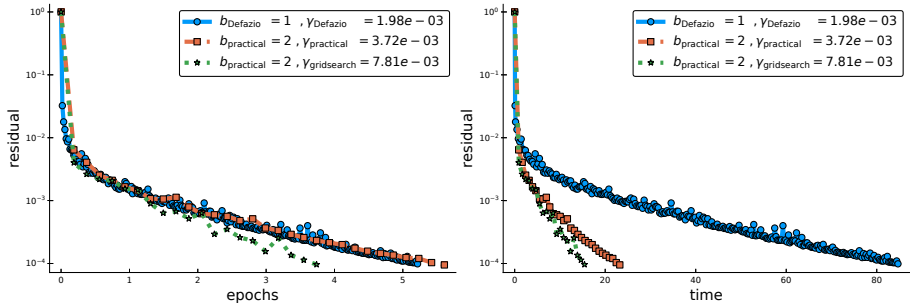


(b) $\lambda = 10^{-3}$

Figure 19: Performance of SAGA implementations for the feature-scaled dataset *slice*.

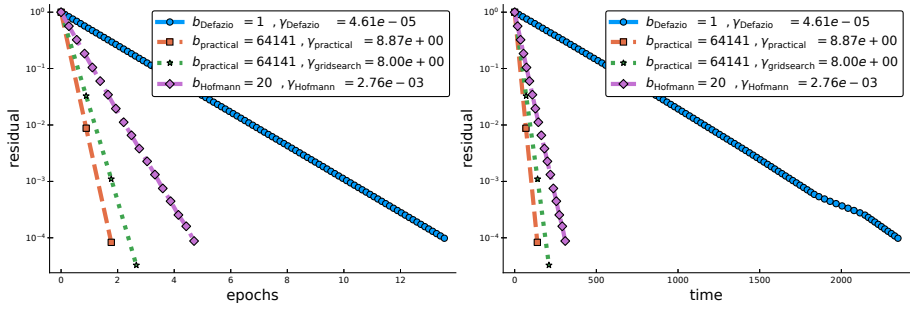


(a) $\lambda = 10^{-1}$

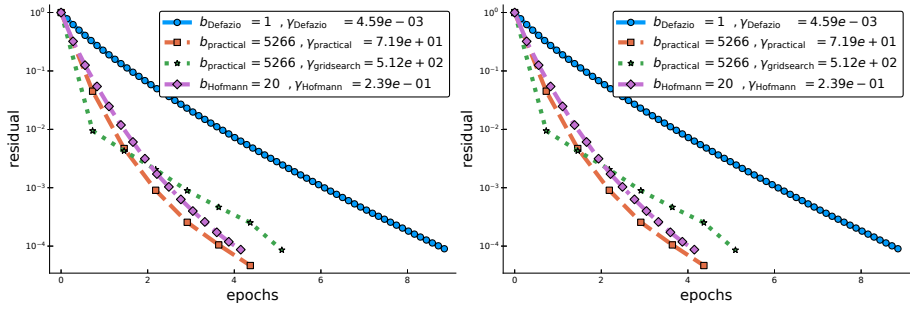


(b) $\lambda = 10^{-3}$

Figure 20: Performance of SAGA implementations for the unscaled dataset *slice*.

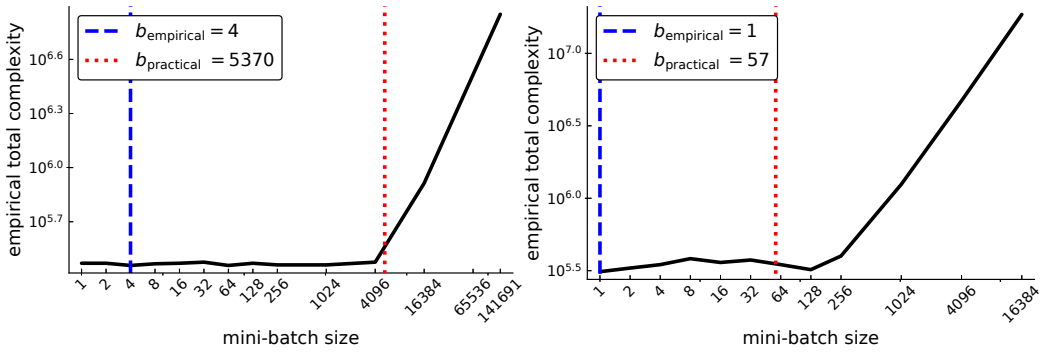


(a) $\lambda = 10^{-1}$



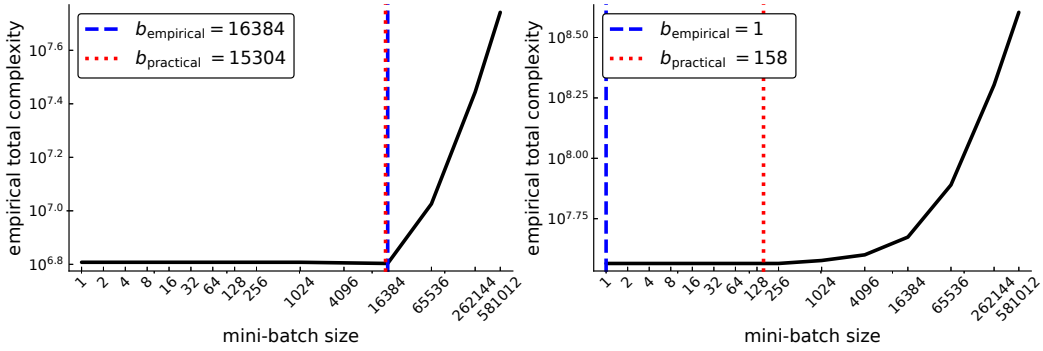
(b) $\lambda = 10^{-3}$

Figure 21: Performance of SAGA implementations for the unscaled dataset *real-sim*.



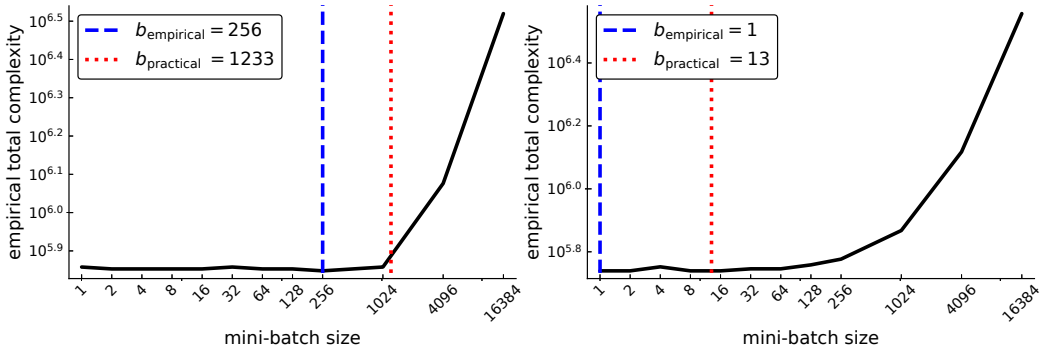
(a) $\lambda = 10^{-1}$ (b) $\lambda = 10^{-3}$

Figure 22: Empirical total complexity versus mini-batch size for the feature-scaled *ijcnn1* dataset.



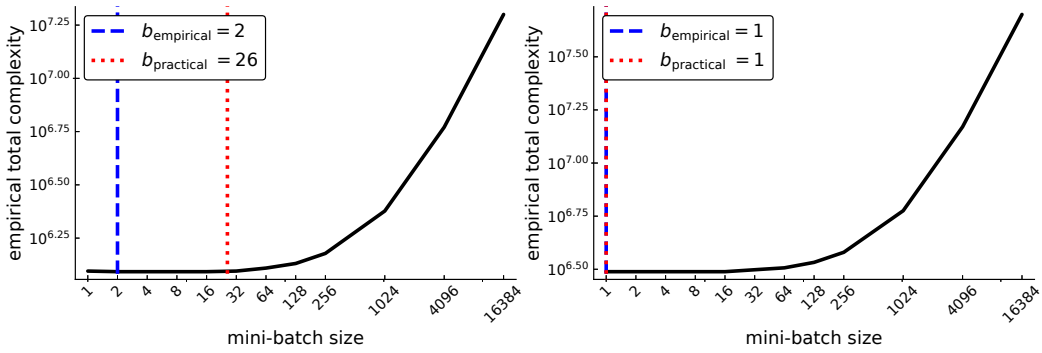
(a) $\lambda = 10^{-1}$ (b) $\lambda = 10^{-3}$

Figure 23: Empirical total complexity versus mini-batch size for the feature-scaled *covtype.binary* dataset.



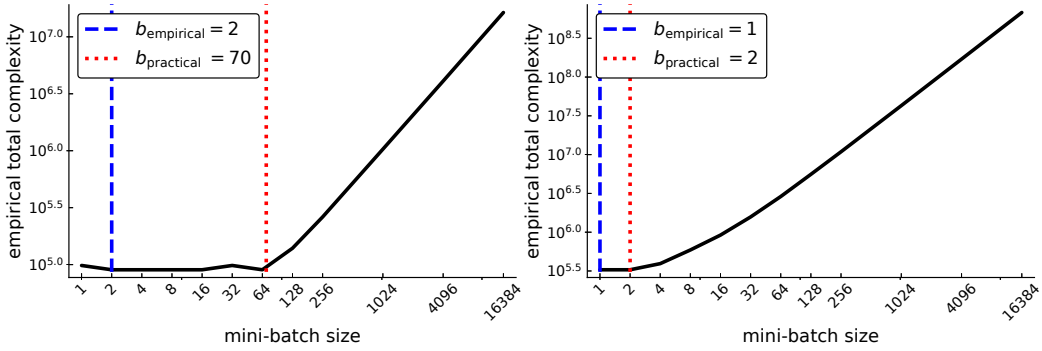
(a) $\lambda = 10^{-1}$ (b) $\lambda = 10^{-3}$

Figure 24: Empirical total complexity versus mini-batch size for the feature-scaled *YearPredictionMSD* dataset.



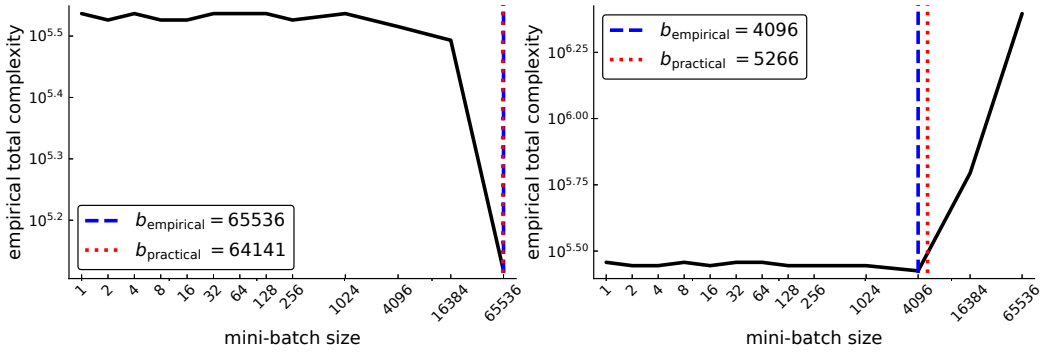
(a) $\lambda = 10^{-1}$ (b) $\lambda = 10^{-3}$

Figure 25: Empirical total complexity versus mini-batch size for the feature-scaled *slice* dataset.



(a) $\lambda = 10^{-1}$ (b) $\lambda = 10^{-3}$

Figure 26: Empirical total complexity versus mini-batch size for the unscaled *slice* dataset.



(a) $\lambda = 10^{-1}$ (b) $\lambda = 10^{-3}$

Figure 27: Empirical total complexity versus mini-batch size for the unscaled *real-sim* dataset.