

A OPA ALGORITHM

Algorithm LBFGS: (Limited memory) BFGS method with OPA

Input: initial guess (z_0, B_0^{-1}) , where B_0^{-1} is symmetric and positive definite, tolerance $\epsilon > 0$, frequency of additional updates $M \in \mathbb{N}$, memory limit $L \in \mathbb{N} \cup \{\infty\}$, (t_n) a null sequence of positive numbers with $\sum_n t_n < \infty$

Let $F := \nabla_z g_\theta$

for $n = 0, 1, 2, \dots$ **do**

if $\|F(z_n)\| \leq \epsilon$ **then** let $z^* := z_n$ and let $B := B_n$; **STOP**

 Let $\hat{B}_n^{-1} := B_n^{-1}$

if $(n \bmod M) = 0$ **then**

 let $e_n := t_n B_n^{-1} \frac{\partial g_\theta}{\partial \theta} \Big|_{z_n}$, $\hat{y}_n := F(z_n + e_n) - F(z_n)$ and $\hat{r}_n := (e_n)^T \hat{y}_n$

if $\hat{r}_n > 0$ **then**

 let $\hat{a}_n := e_n - B_n^{-1} \hat{y}_n$ and let

$$\hat{B}_n^{-1} := B_n^{-1} + \frac{\hat{a}_n (e_n)^T + e_n (\hat{a}_n)^T}{\hat{r}_n} - \frac{(\hat{a}_n)^T \hat{y}_n}{(\hat{r}_n)^2} e_n (e_n)^T$$

 Let $B_n^{-1} := \hat{B}_n^{-1}$

if $n \geq L$ **then** remove update $n - L$ from B_n^{-1}

 Let $p_n := -B_n^{-1} F(z_n)$

 Obtain α_n via line-search and let $s_n := \alpha_n p_n$

 Let $z_{n+1} := z_n + s_n$, $y_n := F(z_{n+1}) - F(z_n)$ and $r_n := (s_n)^T y_n$

if $r_n > 0$ **then**

 let $a_n := s_n - B_n^{-1} y_n$ and let

$$B_{n+1}^{-1} := B_n^{-1} + \frac{a_n (s_n)^T + s_n (a_n)^T}{r_n} - \frac{(a_n)^T y_n}{(r_n)^2} s_n (s_n)^T$$

else let $B_{n+1}^{-1} := B_n^{-1}$

if $n \geq L$ **then** remove update $n - L$ from B_{n+1}^{-1}

Output: z^*, B

Remark. A possible choice for (t_n) is to use an arbitrary $t_0 > 0$ and $t_n := \|s_{n-1}\|$ for $n \geq 1$.

B PROOFS OF SHINE CONVERGENCE

To facilitate reading, we restate the results before proving them.

B.1 CONVERGENCE USING ULI

Theorem 2 (Convergence of SHINE to the Hypergradient using ULI). *Let us denote $p_\theta^{(n)}$, the SHINE direction for iterate n in [Algorithm 1](#) with $b = \text{true}$. Under Assumptions 1 and 2, for a given parameter θ , (z_n) converges q -superlinearly to z^* and*

$$\lim_{n \rightarrow \infty} p_\theta^{(n)} = \frac{\partial \mathcal{L}}{\partial \theta} \Big|_{z^*}$$

Proof. Under Assumptions 1 and 2, [More and Trangenstein \(1976, Theorem 5.7\)](#) shows that B_n satisfies

$$\lim_{n \rightarrow \infty} B_n = J_{g_\theta}(z^*)$$

The inversion operator is continuous in the space of invertible matrices, so we have:

$$\lim_{n \rightarrow \infty} B_n^{-1} = J_{g_\theta}(z^*)^{-1}$$

Because $\nabla_z \mathcal{L}$ and $\frac{\partial g_\theta}{\partial \theta}$ are continuous at z^* by [Assumption 2](#) (iii), we also have thanks to [Assumption 2](#) (i):

$$\lim_{n \rightarrow \infty} \nabla_z \mathcal{L}(z_n) = \nabla_z \mathcal{L}(z^*) \quad \text{and} \quad \lim_{n \rightarrow \infty} \frac{\partial g_\theta}{\partial \theta} \Big|_{z_n} = \frac{\partial g_\theta}{\partial \theta} \Big|_{z^*}$$

By continuity we then deduce that, as claimed,

$$\lim_{n \rightarrow \infty} p_\theta^{(n)} = \lim_{n \rightarrow \infty} \nabla_z \mathcal{L}(z_n) B_n^{-1} \frac{\partial g_\theta}{\partial \theta}(z_n) = \nabla_z \mathcal{L}(z^*) J_{g_\theta}(z^*)^{-1} \frac{\partial g_\theta}{\partial \theta} \Big|_{z^*} = \frac{\partial \mathcal{L}}{\partial \theta} \Big|_{z^*} \quad \square$$

B.2 CONVERGENCE FOR BFGS WITH OPA

Assumption 5 (Extended Assumptions for BFGS). *Let $g_\theta(z) = \nabla_z r_\theta(z)$ for some C^2 function $r_\theta : \mathbb{R}^d \rightarrow \mathbb{R}$. Consider [Algorithm 1](#) with $b = false$ and suppose that*

1. *the set $\Omega := \{z \in \mathbb{R}^d : r_\theta(z) \leq r_\theta(z_0)\}$ is convex;*
2. *r_θ is strongly convex in an open superset of Ω (this implies that r_θ has a unique global minimizer z^*) and has a Lipschitz continuous Hessian near z^* ;*
3. *there are positive constants η_1, η_2 such that the line search used in the algorithm ensures that for each $n \geq 0$ either*

$$r_\theta(z_{n+1}) \leq r_\theta(z_n) - \eta_1 \left[\frac{\nabla r_\theta(z_n)^T p_n}{\|p_n\|} \right]^2 \quad \text{or} \quad r_\theta(z_{n+1}) \leq r_\theta(z_n) + \eta_2 \nabla r_\theta(z_n)^T p_n$$

is satisfied;

4. *the line search has the property that $\alpha_n = 1$ will be used if both*

$$\frac{\|(B_n - J_{g_\theta}(z_n))s_n\|}{\|s_n\|} \quad \text{and} \quad \|z_n - z^*\|$$

are sufficiently small.

Remark. *The requirements 3. and 4. on the line search are, for instance, satisfied under the well-known Wolfe conditions, see [Byrd et al. \(1988, section 3\)](#) for further comments.*

Theorem 3 (Convergence of SHINE to the Hypergradient for BFGS with OPA). *Let us consider $p_\theta^{(n)}$, the SHINE direction for iterate n in [Algorithm 1](#) that is enriched by extra updates in the direction e_n defined in (5). Under Assumptions 2 (ii-iii) and 3, for a given parameter θ , we have the following: [Algorithm 1](#), for any symmetric and positive definite matrix B_0 , generates a sequence (z_n) that converges q -superlinearly to z^* , and there holds*

$$\lim_{n \rightarrow \infty} p_\theta^{(n)} = \frac{\partial \mathcal{L}}{\partial \theta} \Big|_{z^*} \quad (6)$$

Proof. The proof is divided into four steps. The first step is to establish the q -superlinear convergence of (z_n) to z^* . Denoting by $N_e \subset \{0, M, 2M, \dots\}$ the set of indices of extra updates that are actually applied, the second step consists of showing

$$\lim_{N_e \ni n \rightarrow \infty} (B_n - J_{g_\theta}(z^*)) \frac{e_n}{\|e_n\|} = 0, \quad (9)$$

where, in this proof, B_n always represents the matrix from [Algorithm LBFGS](#) before the update in the direction e_n is applied, i.e., the matrix whose inverse appears in the definition of e_n , while \hat{B}_n always represents the matrix from [Algorithm LBFGS](#) after the update in the direction e_n has been applied; if the update in the direction e_n is not applied, then $B_n = \hat{B}_n$. The third step is to prove that (9) implies the desired convergence (6) of the SHINE direction if the limit $n \rightarrow \infty$ is replaced by $N_e \ni n \rightarrow \infty$, i.e., the limit is taken on the subsequence corresponding to N_e . The fourth step is then to transfer the convergence to the entire sequence.

It is easy to check that instead of updating B_n^{-1} , respectively, \hat{B}_n^{-1} , we can also obtain the sequences (B_n) and (\hat{B}_n) by updating according to

$$B_{n+1} = B_n + \frac{y_n y_n^T}{y_n^T s_n} - \frac{B_n s_n (B_n s_n)^T}{s_n^T B_n s_n}$$

for the usual update (skipping the update if $y_n^T s_n \leq 0$), respectively,

$$\hat{B}_n = B_n + \frac{\hat{y}_n \hat{y}_n^T}{\hat{y}_n^T e_n} - \frac{B_n e_n (B_n e_n)^T}{e_n^T B_n e_n}$$

for the extra update (skipping the update if $\hat{y}_n^T e_n \leq 0$). Here, the quantities y_n , \hat{y}_n and e_n are defined as in [Algorithm LBFGS](#). We can now argue essentially as in the proof of [Byrd et al. \(1988, Theorem 3.1\)](#) to show that (z_n) converges q-superlinearly to z^* . As part of that proof we obtain that $\hat{B}_n \neq B_n$ for at least $\lceil 0.5Q \rceil$ of the indices $n = 0, M, 2M, \dots, QM$ for any $Q \in \mathbb{N}$ (namely for all $n \in N_e$ satisfying $n \leq QM$) and that we can apply [Byrd and Nocedal \(1989, Theorem 3.2\)](#), which yields

$$\lim_{n \rightarrow \infty} (\hat{B}_n - J_{g_\theta}(z^*)) \frac{s_n}{\|s_n\|} = 0 \quad \text{and} \quad \lim_{N_e \ni n \rightarrow \infty} (B_n - J_{g_\theta}(z^*)) \frac{e_n}{\|e_n\|} = 0. \quad (10)$$

For the third step, we abbreviate $v_n := \frac{\partial g_\theta}{\partial \theta} \big|_{z_n}$. From the definition of e_n and (10) we infer that

$$0 = \lim_{N_e \ni n \rightarrow \infty} (B_n - J_{g_\theta}(z^*)) \frac{e_n}{\|e_n\|} = \lim_{N_e \ni n \rightarrow \infty} (I - J_{g_\theta}(z^*) B_n^{-1}) \frac{v_n}{\|B_n^{-1} v_n\|}.$$

After multiplication with $J_{g_\theta}(z^*)^{-1}$ this entails

$$\lim_{N_e \ni n \rightarrow \infty} (J_{g_\theta}(z^*)^{-1} - B_n^{-1}) \frac{v_n}{\|B_n^{-1} v_n\|} = 0,$$

which shows that

$$\lim_{N_e \ni n \rightarrow \infty} B_n^{-1} v_n = \lim_{N_e \ni n \rightarrow \infty} J_{g_\theta}(z^*)^{-1} v_n = J_{g_\theta}(z^*)^{-1} \frac{\partial g_\theta}{\partial \theta} \big|_{z^*}$$

by [Assumption 2](#) (iii). Using [Assumption 2](#) (iii) again it follows that

$$\lim_{N_e \ni n \rightarrow \infty} p_\theta^{(n)} = \lim_{N_e \ni n \rightarrow \infty} \nabla_z \mathcal{L}(z_n) B_n^{-1} \frac{\partial g_\theta}{\partial \theta} \bigg|_{z_n} = \nabla_z \mathcal{L}(z^*) J_{g_\theta}(z^*)^{-1} \frac{\partial g_\theta}{\partial \theta} \bigg|_{z^*} = \frac{\partial \mathcal{L}}{\partial \theta} \bigg|_{z^*},$$

concluding the third step. To infer that (6) holds, it suffices to show that $\lim_{N_e \ni n \rightarrow \infty} \|B_n - B_{j_n}\| = 0$ for any sequence $(j_n)_{n \in N_e} \subset \mathbb{N}$ such that $\{j_n, j_n + 1, \dots, n - 1\} \cap N_e = \emptyset$ for all $n \in N_e$ sufficiently large. Indeed, since for $C := \max\{\sup_n \|B_n\|, \sup_n \|B_n^{-1}\|\}$, which is finite by [Byrd and Nocedal \(1989, Theorem 3.2\)](#), there holds

$$(B_n) \subset \left\{ A \in \mathbb{R}^{d \times d} : A^{-1} \text{ exists, } \|A\| \leq C, \|A^{-1}\| \leq C \right\}$$

and the set on the right-hand side of the inclusion is compact by the Banach lemma, inversion is a uniformly continuous operation on this set, hence $\lim_{N_e \ni n \rightarrow \infty} \|B_n^{-1} - B_{j_n}^{-1}\| = 0$, so

$$\lim_{N_e \ni n \rightarrow \infty} \|p_\theta^{(n)} - p_\theta^{(j_n)}\| = 0$$

by continuity, and therefore

$$\lim_{N_e \ni n \rightarrow \infty} p_\theta^{(j_n)} = \lim_{N_e \ni n \rightarrow \infty} p_\theta^{(n)} = \frac{\partial \mathcal{L}}{\partial \theta} \bigg|_{z^*}$$

by the third step, establishing the claim.

It remains to show the validity of $\lim_{N_e \ni n \rightarrow \infty} \|B_n - B_{j_n}\| = 0$ for any sequence $(j_n)_{n \in N_e}$ such that $\{j_n, j_n + 1, \dots, n - 1\} \cap N_e = \emptyset$ for all $n \in N_e$ sufficiently large. Since at least every second extra update is actually carried out, the condition on the intersection implies $n - j_n \leq 2M - 1$ for all these n . Now let $(j_n)_{n \in N_e}$ be any such sequence. Then $B_n - B_{j_n} = \sum_{m=j_n}^{n-1} B_{m+1} - B_m$ is a sum of at most $2M - 1$ BFGS updates in search directions, but contains no extra updates. Hence, the secant conditions $B_{n-l} s_{n-1-l} = y_{n-1-l}$, $l \in \{0, 1, \dots, n - j_n\}$, are satisfied, allowing us to deduce

$$\begin{aligned} \|B_{n-l} - B_{n-l-1}\| &= \frac{\|(B_{n-l} - B_{n-l-1}) s_{n-l-1}\|}{\|s_{n-l-1}\|} \\ &\leq \frac{\|y_{n-l-1} - J_{g_\theta}(z^*) s_{n-l-1}\|}{\|s_{n-l-1}\|} + \frac{\|(B_{n-l-1} - J_{g_\theta}(z^*)) s_{n-l-1}\|}{\|s_{n-l-1}\|} \end{aligned}$$

for all $l \in \{0, 1, \dots, n - j_n - 1\}$. For each of these l , both terms on the right-hand side tend to zero for $N_e \ni n \rightarrow \infty$ (for the second term this follows from the first identity in (10) due to $B_{n-l-1} = \hat{B}_{n-l-1}$). Recalling that $B_n - B_{j_n} = \sum_{m=j_n}^{n-1} B_{m+1} - B_m$ we find $\lim_{N_e \ni n \rightarrow \infty} \|B_n - B_{j_n}\| = 0$, which finishes the fourth step and thus concludes the proof. \square

B.3 CONVERGENCE FOR ADJOINT BROYDEN WITH OPA

Theorem 4 (Convergence of SHINE to the Hypergradient for Adjoint Broyden with OPA). *Let us consider $p_\theta^{(n)}$, the SHINE direction for iterate n in Algorithm 1 with the Adjoint Broyden secant condition (7) and extra update in the direction v_n defined in (8). Under Assumptions 2 and 4, for a given parameter θ , we have q -superlinear convergence of (z_n) to z^* and*

$$\lim_{n \rightarrow \infty} p_\theta^{(n)} = \frac{\partial \mathcal{L}}{\partial \theta} \Big|_{z^*}$$

Proof. Due to Assumption 2, the superlinear convergence of (z_n) follows from Schlenkrich et al. (2010, Theorem 2). The proof of the remaining claim is divided into two cases.

Case 1: Suppose that $\nabla_z \mathcal{L}(z^*) = 0$. By continuity this implies $\lim_{n \rightarrow \infty} \nabla_z \mathcal{L}(z_n) = 0$. Since the sequence $(B_n^{-1} \frac{\partial g_\theta}{\partial \theta} |_{z_n})$ is bounded by Assumption 4, it follows that

$$\lim_{n \rightarrow \infty} p_\theta^{(n)} = \lim_{n \rightarrow \infty} \nabla_z \mathcal{L}(z_n) B_n^{-1} \frac{\partial g_\theta}{\partial \theta} \Big|_{z_n} = 0 = \frac{\partial \mathcal{L}}{\partial \theta} \Big|_{z^*},$$

as claimed.

Case 2: Suppose that $\nabla_z \mathcal{L}(z^*) \neq 0$. By continuity this implies $\nabla_z \mathcal{L}(z_n) \neq 0$ for all sufficiently large $n \in \mathbb{N}$. Let us denote by $N_e \subset \mathbb{N}$ the set of indices of extra updates. We stress that this set is infinite since, by construction, every M -th update is an extra update. We have $v_n \neq 0$ for all sufficiently large $n \in N_e$, hence Schlenkrich et al. (2010, Lemma 3) yields

$$\lim_{N_e \ni n \rightarrow \infty} \frac{\|\nabla_z \mathcal{L}(z_n)(I - B_n^{-1} J_{g_\theta}(z^*))\|}{\|(\nabla_z \mathcal{L}(z_n) B_n^{-1})^T\|} = \lim_{N_e \ni n \rightarrow \infty} \frac{\|(v_n)^T (B_n - J_{g_\theta}(z^*))\|}{\|v_n\|} = 0.$$

This implies

$$\lim_{N_e \ni n \rightarrow \infty} \frac{\|\nabla_z \mathcal{L}(z_n)(J_{g_\theta}(z^*)^{-1} - B_n^{-1})\|}{\|\nabla_z \mathcal{L}(z_n) B_n^{-1}\|} = 0,$$

thus necessarily

$$\lim_{N_e \ni n \rightarrow \infty} \|\nabla_z \mathcal{L}(z_n)(J_{g_\theta}(z^*)^{-1} - B_n^{-1})\| = 0.$$

Since $\lim_{N_e \ni n \rightarrow \infty} \nabla_z \mathcal{L}(z_n) J_{g_\theta}(z^*)^{-1} = \nabla_z \mathcal{L}(z^*) J_{g_\theta}(z^*)^{-1}$ by continuity, we find

$$\lim_{N_e \ni n \rightarrow \infty} \nabla_z \mathcal{L}(z_n) B_n^{-1} = \nabla_z \mathcal{L}(z^*) J_{g_\theta}(z^*)^{-1},$$

whence

$$\lim_{N_e \ni n \rightarrow \infty} p_\theta^{(n)} = \lim_{N_e \ni n \rightarrow \infty} \nabla_z \mathcal{L}(z_n) B_n^{-1} \frac{\partial g_\theta}{\partial \theta} \Big|_{z_n} = \nabla_z \mathcal{L}(z^*) J_{g_\theta}(z^*)^{-1} \frac{\partial g_\theta}{\partial \theta} \Big|_{z^*} = \frac{\partial \mathcal{L}}{\partial \theta} \Big|_{z^*}, \quad (11)$$

where we have used continuity again. To prove that these limits hold not only for $N_e \ni n \rightarrow \infty$ but in fact for all $\mathbb{N} \ni n \rightarrow \infty$, we establish, as intermediate claim, that for any fixed $m \in \mathbb{N}$ we have $\lim_{n \rightarrow \infty} \|B_{n+m} - B_n\| = 0$. Note that this claim is equivalent to $\lim_{n \rightarrow \infty} \|B_{n+1} - B_n\| = 0$. Denoting by $L \geq 0$ the Lipschitz constant of J_{g_θ} near z^* , we find

$$\begin{aligned} \|B_{n+1} - B_n\| &= \frac{\|v_n v_n^T [J_{g_\theta}(z_{n+1}) - B_n]\|}{\|v_n\|^2} \leq \|J_{g_\theta}(z_{n+1}) - J_{g_\theta}(z^*)\| + \frac{\| [J_{g_\theta}(z^*) - B_n]^T v_n \|}{\|v_n\|} \\ &\leq L \|z_{n+1} - z^*\| + \frac{\|E_n^T v_n\|}{\|v_n\|}. \end{aligned}$$

Both terms on the right-hand side go to zero as n goes to infinity: the first one due to $\lim_{n \rightarrow \infty} z_n = z^*$ and the second one since $\lim_{n \rightarrow \infty} \frac{\|E_n^T v_n\|}{\|v_n\|} = 0$ by Schlenkrich et al. (2010, Lemma 3). This shows that $\lim_{n \rightarrow \infty} \|B_{n+1} - B_n\| = 0$, which concludes the proof of the intermediate claim.

From $\lim_{n \rightarrow \infty} \|B_{n+m} - B_n\| = 0$ for any fixed $m \in \mathbb{N}$ it follows that for any sequence $(j_n) \subset \mathbb{N}$ with $\sup_n |j_n - n| < \infty$ there holds $\lim_{n \rightarrow \infty} \|B_{j_n} - B_n\| = 0$. This implies for any such sequence (j_n) the limit $\lim_{n \rightarrow \infty} \|B_{j_n}^{-1} - B_n^{-1}\| = 0$. To establish this, note that for $C := \max\{\sup_n \|B_n\|, \sup_n \|B_n^{-1}\|\}$, which is finite by Assumption 4 and the combination of the bounded deterioration principle (Schlenkrich et al., 2010, Lemma 2) with Assumption 2 (i), the set

$$\left\{ A \in \mathbb{R}^{d \times d} : A^{-1} \text{ exists, } \|A\| \leq C, \|A^{-1}\| \leq C \right\}$$

includes the sequence (B_n) and is compact by the Banach lemma, so inversion is a *uniformly* continuous operation on this set.

Now let us construct a sequence $(j_n) \subset N_e$ by defining, for every $n \in \mathbb{N}$, $j_n := \arg \min_{m \in N_e} |n - m|$. That is, for every n , j_n denotes the member of N_e with the smallest distance to n . It is clear that $|n - j_n| \leq M - 1$ for all n , hence $\lim_{n \rightarrow \infty} \|B_{j_n}^{-1} - B_n^{-1}\| = 0$. Using this and, again, continuity it is easy to see that

$$\lim_{n \rightarrow \infty} \|p_\theta^{(n)} - p_\theta^{(j_n)}\| = 0,$$

which implies by (11) that

$$\lim_{n \rightarrow \infty} p_\theta^{(n)} = \lim_{n \rightarrow \infty} p_\theta^{(j_n)} = \lim_{N_e \ni n \rightarrow \infty} p_\theta^{(n)} = \frac{\partial \mathcal{L}}{\partial \theta} \Big|_{z^*},$$

thereby establishing the claim. \square

Remark. An inspection of the proof reveals that if B_n is never updated in the direction z_n , but only updated in the direction v_n defined in (8), then Assumption 4 can be replaced by the significantly weaker assumption that the sequence $(B_n^{-1} \frac{\partial g_\theta}{\partial \theta} |_{z_n})$ is bounded. The price to pay is that the convergence rate of (z_n) to z^* will be slower (q -linear instead of q -superlinear) since the updates in the direction z_n are critical for ensuring fast convergence of (z_n) to z^* .

C LOGISTIC REGRESSION HYPERPARAMETERS

For both datasets we split the data randomly (with a different seed for each run) between training-validation-test, with the following proportions: 90%-5%-5%. The hyperparameters are the same as in the original HOAG work (Pedregosa, 2016), except:

- We use a memory limitation of 30 updates (not grid-searched) for accelerated methods (Jacobian-Free and SHINE), compared to 10 for the original method. This is because the approximation should be better using more updates. We verified that using 30 updates for the original method does not improve the convergence speed. That number is 60 for OPA.
- We use a smaller exponential decrease of 0.78 (not grid-searched) for the accelerated methods, compared to 0.99 for the original method. This is because in the very long run, the approximation can cause oscillations.

We also use the same setting as Pedregosa (2016) for the Grid and Random Search. Finally, we highlight that warm restart is used for both the inner problem and the Hessian inversion in the direction of the gradient.

OPA inversion experiments For the OPA experiments, we used a memory limitation of 60, and a tolerance of 10^{-6} . The OPA update is done every 5 regular updates.

D DEQ TRAINING DETAILS

The training details are the same as the original Multiscale DEQ paper (Bai et al., 2020): all the hyperparameters are kept the same and not fine-tuned, and the data split is the same. We recall here some important aspects. For both datasets, the network is first trained in an unrolled weight-tied fashion for a few epochs in order to stabilize the training.

We also underline that the DEQ models, in addition to having a fixed-point-defining sub-network, also have a classification and a projection head.

Finally, for Figure 3, the median backward pass is computed with 100 samples on a single V100 GPU for a batch size of 32.

D.1 CIFAR

Adam optimizer (Kingma and Ba, 2015) is used with a 10^{-3} start learning rate, and a cosine annealing schedule.

D.2 IMAGENET

The Stochastic Gradient Descent optimizer is used with a 5×10^{-2} start learning rate, and a cosine annealing schedule.

The images are downsampled 2 times before being fed to the fixed-point defining sub-network.

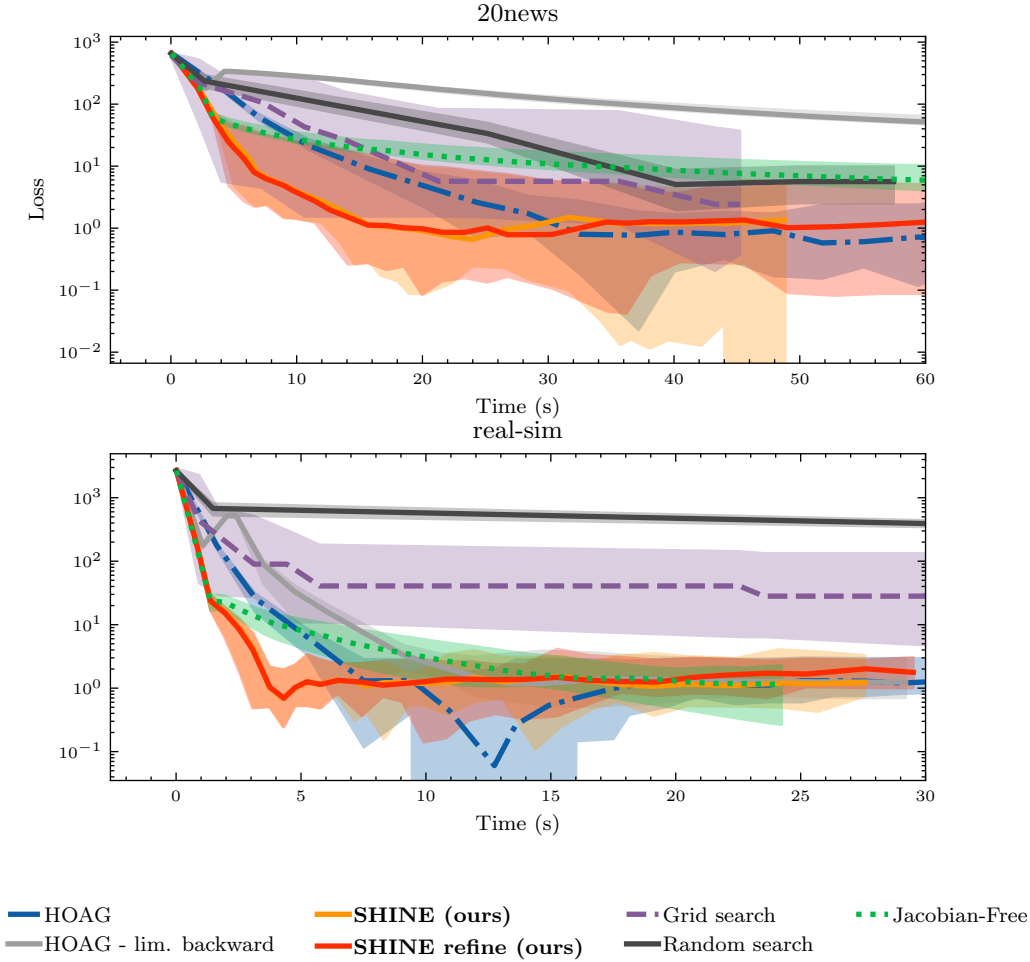


Figure E.1: **Bi-level optimization:** Convergence of different hyperparameter optimization methods on the ℓ_2 -regularized logistic regression problem for two datasets (20news (Lang, 1995) and real-sim (lib)) on held-out test data.

E ADDITIONAL RESULTS

E.1 BI-LEVEL OPTIMIZATION EXTENDED

In order to make sure that SHINE was indeed improving over HOAG (Pedregosa, 2016), we also looked at the results obtained when performing an inversion with a precision lower than that prescribed by Pedregosa (2016) originally (i.e. truncating the iterative inversion). These results, also complemented with Random Search (Bergstra and Bengio, 2012), can be seen in Figure E.1. They confirm that the advantage provided by SHINE cannot be retrieved with a looser tolerance on the inversion.

E.2 CONTRACTIVITY ASSUMPTION

One of the main limiting assumptions in the original Jacobian-Free method work (Fung et al., 2021), is the contractivity assumption. We showed here that it was not important to enforce this in order to achieve excellent results, but one can wonder whether this assumption is not met in practice thanks to the unrolled pretraining of DEQs. We looked at the contractivity of the fixed-point defining sub-network empirically by using the power-method applied to a non-linear function, in the CIFAR setting. The results, summarized in Table E.1, show that the fixed-point defining sub-network is not contractive at all.

Table E.1: Non-linear spectral radius obtained by the power method for the fixed-point defining sub-network for the 3 different methods.

Method	Non-linear spectral radius
Original	230.5
Jacobian-Free	193.7
SHINE	234.2

Table E.2: The time required for each method on the different datasets during the equilibrium training. For the forward and backward passes, the time is measured offline, for a single batch of 32 samples, with a single GPU, using the median to avoid outliers. This time is given in milliseconds. For the epochs, the time is measured by taking an average of the 6 first epochs, and given in hours-minutes for Imagenet and minutes-seconds for CIFAR. The epoch time for SHINE without improvement on Imagenet is not given because it never reaches the 26 forward steps: the implicit depth is too short. Fallback is not used for CIFAR. Numbers in parenthesis indicate the number of inversion steps for the refined versions.

Dataset Name	CIFAR (Krizhevsky, 2009)			ImageNet (Deng et al., 2009)		
Method Name	Forward	Backward	Epoch	Forward	Backward	Epoch
Original (Bai et al., 2020)	256	210	4min40	644	798	3h38
Jacobian-Free (Fung et al., 2021)	249	12.9	3min10	621	13.5	2h02
SHINE Fallback (ours)	218	16.0	3min20	622	35.3	2h13
SHINE Fallback refine (5, ours)	272	96.6	3min50	622	212	2h44
Jacobian-Free refine (5)	260	86.5	3min40	620	186	2h43
Original limited backprop	281	86.4	3min50	653	187	2h40

E.3 TIME GAINS

Because the total training time is not only driven by backward pass but also by the forward pass and the evaluation, we show for completeness in Table E.2 the time gains for the different acceleration methods for the overall epoch. We do not report in this table the time taken for pre-training which is equivalent across all methods, and is not something on which SHINE has an impact. It is clear in Table E.2 that accelerated methods can have a significant impact on the training of DEQs because we see that half the time of the total pass is spent on the backward pass (more on ImageNet (Deng et al., 2009)). We also notice that while SHINE has a slightly slower backward pass than the Jacobian-Free method (Fung et al., 2021), the difference is negligible when compared to the total pass computational cost.

E.4 DEQ OPA RESULTS

We can clearly see in Figure E.2 that in the case of DEQs, OPA also significantly improves the inversion over the other accelerated methods. We also see that the improvements of SHINE over the Jacobian-Free method without OPA are marginal.

Because the inversion is so good, we would expect that the performance of SHINE with OPA would be on par with the original method's. However, this is not what we see in the results presented in Table E.3. Indeed, OPA does improve on SHINE with only Adjoint Broyden, but it does not outperform SHINE done with Broyden.

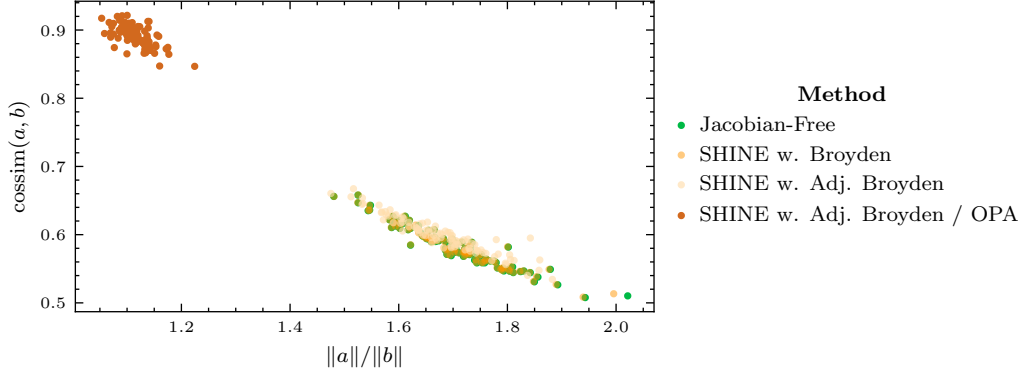


Figure E.2: **Quality of the inversion using OPA in DEQs :** Ratio of the inverse approximation over the exact inverse function of the cosine similarity between the inverse approximation $b = \nabla_z \mathcal{L}(z^*) B_n^{-1}$ and the exact inverse $a = \nabla_z \mathcal{L}(z^*) J_{g_\theta}(z^*)^{-1}$ for different methods. For OPA, the extra update frequency is 5. 100 runs were performed with different batches.

Table E.3: **CIFAR DEQ OPA results :** Top-1 accuracy of different methods on the CIFAR dataset, and epoch mean time.

Method name	Top-1 Accuracy (%)	Epoch mean time
Original	93.51	4min40
Jacobian-Free	93.09	3min10
SHINE (Broyden)	93.14	3min20
SHINE (Adj. Broyden)	92.89	4min
SHINE (Adj. Broyden/OPA)	93.04	4min40