Supplementary Materials for the Paper "CWY Parametrization: a Solution for Parallelized Optimization of Orthogonal and Stiefel Matrices"

In this Appendix we provide the following:

- In Section A, Stiefel RGD through the Sherman-Morrison-Woodbury formula
- In Section B, Hidden state gradients of ConvNERU
- In Section C, Copying, pixel-by-pixel MNIST and time comparison details
- In Section D, Neural machine translation details
- In Section E, Video prediction details
- In Section F, Proofs of results

A STIEFEL RGD THROUGH THE SHERMAN-MORRISON-WOODBURY FORMULA

RGD with Cayley retraction requires inverting the $N \times N$-sized matrix $\eta_k A^{(k-1)}$ thus becoming cubic in N. To make the computation more tractable, Tagare (2011) proposes to use the Sherman-Morrison-Woodbury formula which reduces the size of the inverted matrix to $2M \times 2M$ when the canonical inner product is chosen for RGD. A straightforward extension of Tagare's approach to the Euclidean inner product would require to invert $3M \times 3M$-sized matrix. To demonstrate that, we adapt derivations of Tagare (2011) for canonical inner product and extend them to Euclidean inner product. The following Lemma shows how to compute update $g_k(\eta_k)$ in time $O(NM^2 + M^3)$ without constructing $A^{(k-1)}$ explicitly.

Lemma 1. *Consider $\Omega \in \mathbb{R}^{N \times M}$ and $A = BC^\top \in Skew(N)$ for some matrices $B, C \in \mathbb{R}^{N \times D}$, $D \leq N$. Then*

$$\mathit{Cayley}(A)\Omega = \Omega - B\left(I + \frac{1}{2}C^\top B\right)^{-1}\left(C^\top \Omega\right) \tag{4}$$

Proof. We first need to show that the right hand side of (4) always exists, i.e. $I + \frac{1}{2}C^\top B$ is nonsingular:

$$\det(I + \frac{1}{2}C^\top B) = \det(I + \frac{1}{2}BC^\top) = \det(I + \frac{1}{2}A) \neq 0$$

where in the first transition we apply Sylvester's determinant identity. $I + \frac{1}{2}A$ is nonsingular, because the spectrum of any skew-symmetric matrix is pure-imaginary (Theorem 12.9 from Gallier, 2011). So the right hand side is well defined.

Through the application of Sherman-Morrison-Woodbury formula we deduce that

$$\begin{aligned}\text{Cayley}(A)\Omega &= \left(I + \frac{1}{2}BC^\top\right)^{-1}\left(I - \frac{1}{2}BC^\top\right)\Omega \\ &= \left(I - \frac{1}{2}B(I + \frac{1}{2}C^\top B)^{-1}C^\top\right)\left(I - \frac{1}{2}BC^\top\right)\Omega \\ &= \Omega - \frac{1}{2}B\left((I + \frac{1}{2}C^\top B)^{-1}(I - \frac{1}{2}C^\top B) + I\right)C^\top\Omega \\ &= \Omega - \frac{1}{2}B(I + \frac{1}{2}C^\top B)^{-1}(2I - C^\top B + C^\top B)C^\top\Omega \\ &= \Omega - B\left(I + \frac{1}{2}C^\top B\right)^{-1}\left(C^\top\Omega\right)\end{aligned}$$

which concludes the proof. □

For convenience denote $\mathcal{G}^{(k-1)} = \frac{\partial f}{\partial \Omega}(\Omega^{(k-1)})$. Depending on the inner product choice we get the following cases:

1. **Canonical inner product**. Then

$$\eta_k A^{(k-1)} = \eta_k \mathcal{G}^{(k-1)} \Omega^{(k-1)\top} - \eta_k \Omega^{(k-1)} \mathcal{G}^{(k-1)\top} = BC^\top$$

where

$$B = \eta_k \begin{bmatrix} \mathcal{G}^{(k-1)} & \Omega^{(k-1)} \end{bmatrix}, \quad C = \begin{bmatrix} \Omega^{(k-1)} & -\mathcal{G}^{(k-1)} \end{bmatrix}, \quad B, C \in \mathbb{R}^{N \times 2M}.$$

2. **Euclidean inner product**. Then

$$\eta_k A^{(k-1)} = \eta_k \mathcal{G}^{(k-1)} \Omega^{(k-1)\top} - \eta_k \Omega^{(k-1)} \mathcal{G}^{(k-1)\top} + \frac{\eta_k}{2} \Omega^{(k-1)} E \Omega^{(k-1)\top} = BC^\top$$

where

$$E = \mathcal{G}^{(k-1)\top} \Omega^{(k-1)} - \Omega^{(k-1)\top} \mathcal{G}^{(k-1)}, \quad B = \eta_k \begin{bmatrix} \mathcal{G}^{(k-1)} & \Omega^{(k-1)} & \frac{1}{2} \Omega^{(k-1)} E \end{bmatrix},$$
$$C = \begin{bmatrix} \Omega^{(k-1)} & -\mathcal{G}^{(k-1)} & \Omega^{(k-1)} \end{bmatrix}, \quad B, C \in \mathbb{R}^{N \times 3M}.$$

B HIDDEN STATE GRADIENTS OF CONVNERU

The convolution operation can be expressed as

$$(\mathcal{K} * G^{(t-1)})_{i,j} = \widehat{\mathcal{K}}^\top \overline{G}^{(t-1)}_{i,j}, \quad \overline{G}^{(t-1)} \in \mathbb{R}^{h \times w \times q^2 f_{out}},$$
$$\overline{G}^{(t-1)}_{i,j} = \mathrm{concat}\left(\{ G^{(t-1)}_{l,p} \,|\, i - \frac{q-1}{2} \leq l \leq i + \frac{q-1}{2}, j - \frac{q-1}{2} \leq p \leq j + \frac{q-1}{2} \} \right)$$

where $G^{(t-1)}_{l,p} \in \mathbb{R}^{f_{out}}$ is a zero vector when l, p are pointing outside image borders (*zero padding*). By definition of $\overline{G}^{(t-1)}$ and $\mathcal{K} * G^{(t-1)}$ we have the following chain of inequalities between Frobenius norms $\| \cdot \|_F$:

$$\|\mathcal{K} * G^{(t-1)})\|_F^2 = \sum_{i,j} \|(\mathcal{K} * G^{(t-1)}))_{i,j}\|_2^2 = \sum_{i,j} \|\widehat{\mathcal{K}}^\top \overline{G}^{(t-1)}_{i,j}\|_2^2 \leq \sum_{i,j} \|\widehat{\mathcal{K}}\|_2^2 \|\overline{G}^{(t-1)}_{i,j}\|_2^2$$
$$= \|\widehat{\mathcal{K}}\|_2^2 \cdot \|\overline{G}^{(t-1)}\|_F^2 \leq q^2 \|\widehat{\mathcal{K}}\|_2^2 \cdot \|G^{(t-1)}\|_F^2$$

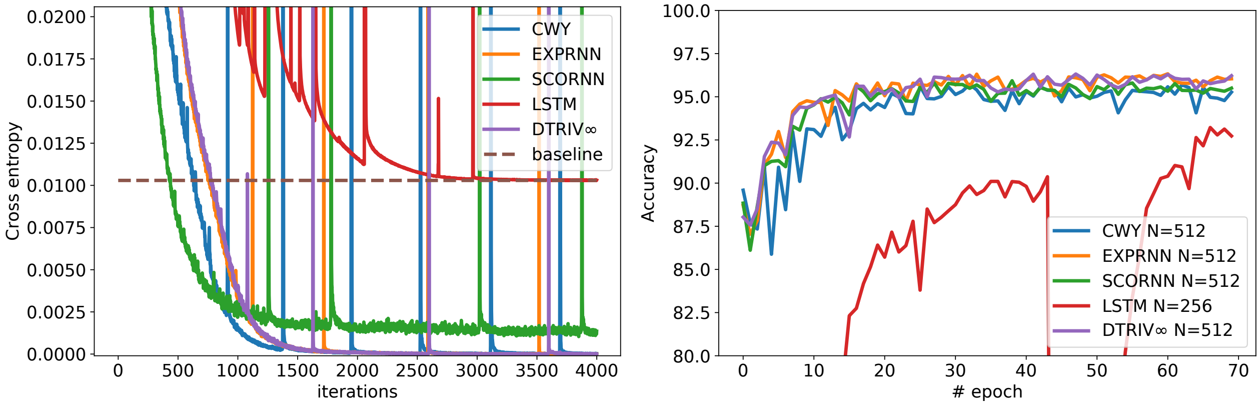
Assuming that $|\sigma(x)| \leq |x|$ which holds for most popular choices of nonlinearity (ReLU, LeakyReLU, tanh), the norm of $G^{(t)}$ cannot grow in exponential manner. The same holds for a sequence of gradients with respect to $\{G^{(t)}\}$, since it is obtained by sequentially applying a transposed linear operator corresponding to "$\mathcal{K}*$" convolution operation and transposition preserves the linear operator norm. This justifies the property of ConvNERU being robust to gradient explosion while allowing long-term information propagation thank to Stiefel convolution kernel. The conducted analysis is reminiscent of Lipschitz constant estimate for image classification CNNs performed by Cisse et al. (2017).

C COPYING, PIXEL-BY-PIXEL MNIST AND TIME COMPARISON: MORE DETAILS AND RESULTS

Results for Copying task ($\mathcal{T} = 2000$), and permuted MNIST (i.e. when pixels in a flatten image are permuted randomly) are shown on Figure 4. For all setups but SCORNN in the Copying task we used initialization technique from (Henaff et al., 2016), whilst for SCORNN we used initialization from (Helfrich et al., 2018). For all setups in the Pixel-by-pixel MNIST (whether permuted or not) we used initialization from (Helfrich et al., 2018). While our results on Pixel-by-pixel MNIST match those of Mhammedi et al. (2017), Mhammedi et al. (2017) were not able to provide comparable results for the Copying task. We observe that correct initialization is crucial for this task.

To initialize CWY, we, first of all, initialize a skew-symmetric matrix, as discussed above. Then we take exponent of this matrix, obtaining an orthogonal matrix. Then, in order to initialize vectors $v^{(1)}, \dots, v^{(N)}$, we run the same procedure as in the Theorem 1 proof (QR decomposition using Householder reflections).

To do the time comparison, we draw elements of $v^{(1)}, \ldots, v^{(N)}$ for CWY from a standard normal distribution. For matrix exponent and Cayley map, we initialize skew symmetric arguments as $\mathcal{X} - \mathcal{X}^\top$, where entries of $\mathcal{X}$ are sampled from a standard normal distribution.



(a)

(b)

Figure 4: **(a)** Copying task, $\mathcal{T} = 2000$. **(b)** Permuted Pixel-by-pixel MNIST, test accuracy.

D NEURAL MACHINE TRANSLATION: MORE DETAILS AND RESULTS

We take aligned bi-texts between the source and target languages and, as preprocessing, remove accents and return word pairs in the form [English, Spanish]. The resulting dataset has an average sequence length of ≈ 17 for both the input and target sequences.

Using a single Tensor Processing Unit (TPU) per model, we train several models from scratch, with no pre-training, on 80,000+ sentence pairs and test on the remaining 20,000+ pairs from the full 100,000+ pair dataset to compare their learning capabilities and stability. We use JAX[1] library for the implementation. Given that we evaluate all models on the same corpus and that our goal is to benchmark across architectures, we elected to employ no pre-training and examine/compare cross-entropy loss directly.

See Figure 5 for the architecture illustration. In our experiments, we used a batch size of 64, an embedding dimension size of 256, and a learning rate of 10^{-2}. For hyperparameter sweeps, we ran experiments with smaller hidden unit sizes. We also experimented with larger and smaller learning rates. Ultimately, for simplicity and clarity, we only present results using the parameters described above.

For additional experimental results, see Table 5.

E VIDEO PREDICTION: MORE DETAILS

All videos, 4 seconds in average, are recorded with a static camera with 25 fps frame rate and frame size of 160 × 120 pixels. We crop and resize each frame into 128 × 128 pixels and then reshape each frame into $64 \times 64 \times 4$ by moving groups of 2×2 pixels into channel dimension. Since each video sequence has a different number of frames, we employ zero padding during batch construction. We use persons with indices 1-12 for training, 13-16 for validation and 17-25 for testing. See Table 6 for KTH dataset statistics.

Given a sequence of known frames $\mathcal{I}^{(1)}, \ldots, \mathcal{I}^{(t)} \in [0,1]^{64\times64\times4}$, the network outputs a prediction $\widehat{\mathcal{I}}^{(t+1)}$ of the next frame $\mathcal{I}^{(t+1)}$. The network is designed as a recurrent block composed of several convolutional recurrent units stacked together with the sequence $\{\mathcal{I}^{(i)}\}_{i=1}^{t}$ passed to the input. In order to increase the receptive field of the recurrent architecture while maintaining a tractable training procedure, we adapt a simplified version of the video prediction architecture from Lee et al. (2018); Ebert et al. (2017). Namely, we stack several recurrent units with a bottleneck structure (hidden sizes $32\times32\times32 \to 16\times16\times64 \to 8\times8\times128 \to 16\times16\times64 \to 32\times32\times32$) and skip connections. We alternate recurrent layers with strided convolutions and then deconvolutions. After

[1] https://jax.readthedocs.io/en/latest/

Table 5: Tatoeba Spa-to-Eng NMT results. We ran 3 seeds for each model. Below we present the average test loss across these seeds, as well as the associated standard deviation. We did not run additional seeds for non-CWY orthogonal parameterization approaches as these methods are slow (requiring many TPU hours to train) and our primary comparison with them was w.r.t. speed.

MODEL	**TEST CE LOSS**	**STANDARD ERROR**
RNN	0.74	.08
GRU	0.56	.05
LSTM	0.55	.05
RGD	2.01	.14
CWY, $L = 1024$	0.56	.03
CWY, $L = 512$	0.66	.03
CWY, $L = 256$	0.64	.06
CWY, $L = 128$	0.50	.01
CWY, $L = 64$	0.60	.01

Table 6: KTH action dataset statistics.

STATISTIC	**WALK**	**JOG**	**RUN**	**BOX**	**WAVE**	**CLAP**
Min sequence length	62	42	26	42	62	24
Max sequence length	231	152	111	362	245	235
Mean sequence length	109.3	68.0	48.9	110.3	129.0	106.2
Total frames count (train set)	20122	12730	9096	20515	23958	19529
Total frames count (val. set)	7622	4551	3448	7558	8436	6415
Total frames count (test set)	15991	9913	7018	15277	18963	16125

each convolution and deconvolution we place a ReLU nonlinearity, as well as using ReLU as the recurrent nonlinearity σ. In the proposed architecture a prediction $\widehat{\mathcal{I}}^{(t+1)}$ is conditioned upon $\mathcal{I}^{(t)}$ through bottleneck and skip connections and conditioned upon $\{\mathcal{I}^{(t')}\}_{t'<t}$ through recurrent temporal connections. See Figure 6 for architecture illustration.

We opt for batch size of 3, recurrent kernel size $q = 3$, learning rate of 10^{-3}. Our experiments are implemented in Tensorflow and run on a single Nvidia Tesla P100 GPU for each experiment. For each experiment we run 150 epochs and choose the model's state showing smallest validation loss value for testing.

F PROOFS

F.1 Theorem 1

Proof. The proof proceeds by induction in N. For $N = 1$ such Q is unique and is equal to $\begin{bmatrix}-1\end{bmatrix}$. So simply take $u_1 = \begin{bmatrix}-1\end{bmatrix}$. Now assume the statement is true for $N = k - 1 \geq 1$. When $N = k > 1$ we consider Q's first column $q = \begin{bmatrix}q_1 & \dots & q_N\end{bmatrix}^\top$ and define a vector $v \in \mathbb{R}^k$ as follows:

$$v = \begin{cases} \frac{q - e^{(1)}}{\|q - e^{(1)}\|} & \text{if } |q_1| < 1 \\ \begin{bmatrix}0 & \dots & 0 & 1\end{bmatrix}^\top & \text{if } q_1 = 1 \\ e^{(1)} & \text{if } q_1 = -1 \end{cases} \tag{5}$$

Observe that

$$H(v)Q = \begin{bmatrix}1 & r^\top \\ \mathbf{0} & Q'\end{bmatrix} \tag{6}$$

for some $r \in \mathbb{R}^{k-1}$. From the fact that $H(v)Q \in \mathcal{O}(k)$ we deduce:

$$\begin{bmatrix}1 & \mathbf{0} \\ r & Q'^\top\end{bmatrix} \begin{bmatrix}1 & r^\top \\ \mathbf{0} & Q'\end{bmatrix} = \begin{bmatrix}1 & r^\top \\ r & Q'^\top Q' + rr^\top\end{bmatrix} = I \tag{7}$$

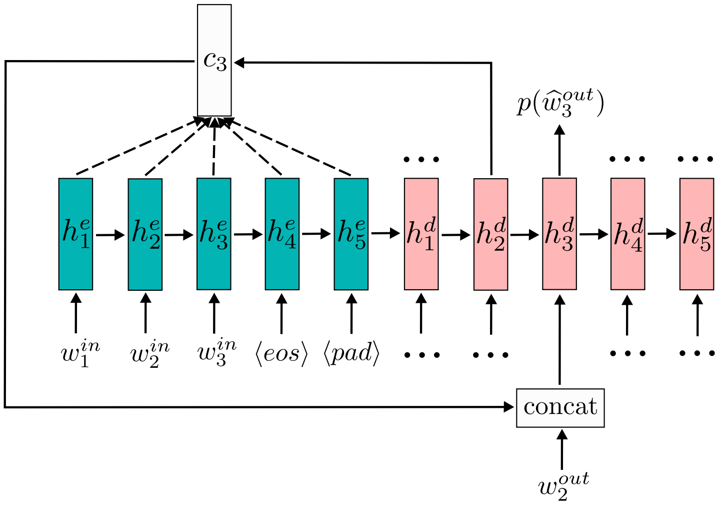


Figure 5: Sketch of the architecture used for Neural Machine Translation experiments. For ease of illustration we use 5 as maximal input and output length. w_i^{in}, w_i^{out} are input and output word embeddings respectively, $\langle eos \rangle$ and $\langle pad \rangle$ denote embeddings of "end of sentence" and "padding" tag respectively. We use two different RNN units for the encoder rollout $h_1^e \to \cdots \to h_5^e$ (blue) and decoder rollout $h_1^d \to \cdots \to h_5^d$ (pink). We illustrate how the distribution of predicted output word $\widehat{w}_3^{out}$ is computed, other output words are processed similarly. Given h_2^d, the context vector $c_3 \in \mathbb{R}^N$ is computed as $\sum_i \alpha_i h_i^e$ where $\sum_i \alpha_i = 1$, $\alpha_i \propto \exp(v^\top \tanh(W_1 h_i^e + W_2 h_2^d))$, $v \in \mathbb{R}^N, W_1, W_2 \in \mathbb{R}^{N \times N}$ are learnable parameters. Then c_3 is concatenated with previous word embedding (w_2^{out} or null tag embedding for the first predicted word) and passed into decoder RNN as input. Decoder RNN output (h_3^d) is passed through linear layer + softmax to obtain a distribution over $\widehat{w}_3^{out}$.

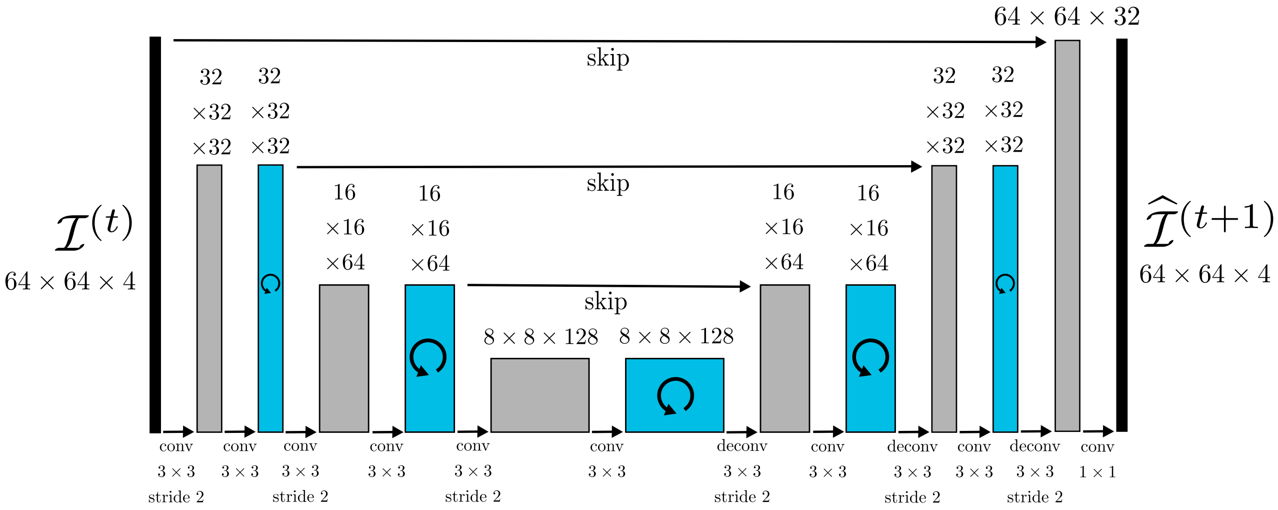


Figure 6: Sketch of the architecture used for video prediction experiments. Blue and grey blocks illustrate hidden representations with and without recurrent connections respectively. We compare different designs of the blue block (ConvLSTM, ConvNERU). In our comparison we try different designs of blue recurrent units with everything else unchanged. As in the original papers (Lee et al., 2018; Ebert et al., 2017), we find that ConvLSTM version works best when instance normalization (Ulyanov et al., 2016) is added after each convolution and before the nonlinearity, including convolutions inside ConvLSTM. We don't use instance normalization with other model variants.

Hence, $r = \mathbf{0}$ and $Q' \in \mathcal{O}(k-1)$. By Sylvester determinant identity $\det(I - 2vv^\top) = 1 - 2v^\top v = -1$, therefore $\det Q' = (-1)^{k-1}$. By the induction step assumption there exist nonzero $v'^{(1)}, \dots, v'^{(k-1)} \in \mathbb{R}^{k-1}$ s.t.

$$Q' = H(v'^{(1)}) \dots H(v'^{(k-1)})$$

We define $v^{(2)} = \begin{bmatrix} 0 & v'^{(1)\top} \end{bmatrix}^\top, \dots, v^{(k)} = \begin{bmatrix} 0 & v'^{(k-1)\top} \end{bmatrix}^\top$ and obtain that

$$H(v)Q = H(v^{(2)}) \dots H(v^{(k)}) \tag{8}$$

Finally, we define $v^{(1)} = v$, left-multiply (8) by $H(v^{(1)})$ and complete the induction step. □

F.2 Theorem 2

Proof. First, observe that S is upper-triangular matrix with $\frac{1}{2}$ on the diagonal. Hence, it is nonsingular and the Theorem statement is valid. Now the proof proceeds by induction in L. For $L = 1$ Theorem is trivial. Suppose Theorem is true for $L = k - 1 \geq 1$. Then the following is true:

$$H(v^{(1)}) \dots H(v^{(k-1)}) = I - U'S'^{-1}U'^\top$$

where $U' = \begin{bmatrix} \frac{v^{(1)}}{\|v^{(1)}\|_2} & \dots \frac{v^{(k-1)}}{\|v^{(k-1)}\|_2} \end{bmatrix}$ and

$$S' = \frac{1}{2}I + \mathrm{striu}(U'^\top U')$$

Then for $L = k$ we get:

$$\begin{aligned} H(v^{(1)}) \dots H(v^{(k)}) &= (I - U'S'^{-1}U'^\top)H(v^{(k)}) \\ &= I - U'S'^{-1}U'^\top - 2\frac{v^{(k)}v^{(k)\top}}{\|v^{(k)}\|_2^2} + 2U'S'^{-1}U'^\top \frac{v^{(k)}v^{(k)\top}}{\|v^{(k)}\|_2^2} \\ &= I - U \begin{bmatrix} S'^{-1} & -2S'^{-1}U'^\top \frac{v^{(k)}}{\|v^{(k)}\|_2^2} \\ \mathbf{0} & 2 \end{bmatrix} U^\top \end{aligned}$$

And the step of induction is completed by observing that

$$\begin{bmatrix} S'^{-1} & -2S'^{-1}U'^\top \frac{v^{(k)}}{\|v^{(k)}\|_2^2} \\ \mathbf{0} & 2 \end{bmatrix} \times S = \begin{bmatrix} S'^{-1} & -2S'^{-1}U'^\top \frac{v^{(k)}}{\|v^{(k)}\|_2^2} \\ \mathbf{0} & 2 \end{bmatrix} \times \begin{bmatrix} S' & U'^\top \frac{v^{(k)}}{\|v^{(k)}\|_2^2} \\ \mathbf{0} & \frac{1}{2} \end{bmatrix} = I$$

□

F.3 Theorem 3

Proof. Similarly to Theorem 2, observe that S is upper-triangular matrix with $\frac{1}{2}$ on the diagonal. Hence, it is nonsingular and Theorem's statement is valid.

Observe that for any nonzero vectors $v^{(1)}, \dots v^{(M)} \in \mathbb{R}^N$

$$\begin{aligned} \left(\begin{bmatrix} I \\ \mathbf{0} \end{bmatrix} - US^{-1}U_1^\top \right)^\top \left(\begin{bmatrix} I \\ \mathbf{0} \end{bmatrix} - US^{-1}U_1^\top \right) &= I + U_1 \Big(S^{-\top}U^\top US^{-1} - S^{-1} - S^{-\top} \Big) U_1^\top \\ &= I + U_1 S^{-\top} \Big(U^\top U - S^\top - S \Big) S^{-1} U_1^\top = I \end{aligned}$$

Hence, $\gamma_{N,M}(v^{(1)}, \dots v^{(M)}) \in \mathrm{St}(N, M)$. To show surjectivity of $\gamma_{N,M}$, consider arbitarary $\Omega \in \mathrm{St}(N, M)$. Let $q = \begin{bmatrix} q_1 & \dots & q_N \end{bmatrix}^\top$ be Ω's first column. We consider value v defined by (5). Using derivations similar to (6-7), we obtain:

$$H(v)\Omega = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \Omega' \end{bmatrix}$$

where $\Omega' \in \mathrm{St}(N-1, M-1)$.

Set $v^{(1)} = v$. Analogously find v' for Ω' such that

$$H(v')\Omega' = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \Omega'' \end{bmatrix}$$

and set $v^{(2)} = \begin{bmatrix} 0 & v'^{\top} \end{bmatrix}^{\top}$. Repeat this procedure $M - 2$ more times to obtain:

$$H(v^{(M)}) \dots H(v^{(1)})\Omega = \begin{bmatrix} I \\ \mathbf{0} \end{bmatrix} \tag{9}$$

Left-multiply (9) by $H(v^{(1)}) \dots H(v^{(M)})$:

$$\Omega = H(v^{(1)}) \dots H(v^{(M)}) \begin{bmatrix} I \\ \mathbf{0} \end{bmatrix}$$

Finally, apply Theorem 2 for series of Householder reflections $H(v^{(1)}) \dots H(v^{(M)})$:

$$\Omega = \left(I - US^{-1}U^{\top} \right) \begin{bmatrix} I \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} I \\ \mathbf{0} \end{bmatrix} - US^{-1}U_1^{\top} = \gamma_{N,M}(v^{(1)}, \dots, v^{(M)})$$

which justifies surjectivity of $\gamma_{N,M}$. □

F.4 Theorem 4

Before providing results which build to the complete proof, we first give a **high-level sketch** to aid intuition. Lemma 2 shows that, for any iteration of SGD, $v^{(1)}, \dots, v^{(L)}$ stay in a region $\mathcal{S} = \{x \in \mathbb{R}^N \mid \|x\|_2 > A\}$, where $A > 0$ is some fixed number. Lemma 3 shows that the composition of f and CWY has Lipschitz-continuous gradients in $\mathcal{S}$. Next, Lemma 4 shows that the gradient proxy has bounded variance in $\mathcal{S}$. The proof itself is essentially Theorem 4.10 from (Bottou et al., 2016), which uses Lipschitz continuity and boundedness to establish SGD convergence guarantees.

Lemma 2. *Suppose conditions of Theorem 4 hold. Since all $v^{(0,1)}, \dots, v^{(0,L)}$ are nonzero, there exists a number $A > 0$ such that for all $l \in \{1, \dots, L\} : A < \|v^{(0,l)}\|_2$. Define a set $\mathcal{S} = \{x \in \mathbb{R}^N \mid \|x\|_2 > A\}$. Then for each $k \geq 0$ $v^{(k,1)}, \dots, v^{(k,L)}$ are well-defined and lie in $\mathcal{S}$.*

Proof. The statement is true for $k = 0$. Suppose it's true for $k - 1$. Since $v^{(k-1,1)}, \dots, v^{(k-1,L)}$ are nonzero, $v^{(k,1)}, \dots, v^{(k,L)}$ are well-defined. Fix $l \in \{1, \dots, L\}$. Observe that for any nonzero $v \in \mathbb{R}^N$: $H(v) = H(\frac{v}{\|v\|_2})$. Hence, $\widetilde{f}(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}))$ can be represented as a function $g(\frac{v^{(k-1,l)}}{\|v^{(k-1,l)}\|_2})$ so that

$$\nabla_{v^{(k-1,l)}} \widetilde{f}(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)})) = \nabla_{v^{(k-1,l)}} g(\frac{v^{(k-1,l)}}{\|v^{(k-1,l)}\|_2}).$$

Denote $s(v) = \frac{v}{\|v\|_2}$. Then

$$\nabla_v g(s(v)) = \frac{1}{\|v\|_2}(I - s(v)s(v)^{\top})\nabla_s g(s(v))$$

and, hence, $v^{\top}\nabla_v g(\frac{v}{\|v\|_2}) = 0$. We use it to derive that for any $\eta \in \mathbb{R}$

$$\begin{aligned} \|v^{(k-1,l)} - \eta\nabla_{v^{(k-1,l)}} g(\frac{v^{(k-1,l)}}{\|v^{(k-1,l)}\|_2})\|_2^2 &= \|v^{(k-1,l)}\|_2^2 + \|\eta\nabla_{v^{(k-1,l)}} g(\frac{v^{(k-1,l)}}{\|v^{(k-1,l)}\|_2})\|_2^2 - 2\eta v^{(k-1,l)\top} g(\frac{v^{(k-1,l)}}{\|v^{(k-1,l)}\|_2}) \\ &= \|v^{(k-1,l)}\|_2^2 + \|\eta\nabla_{v^{(k-1,l)}} g(\frac{v^{(k-1,l)}}{\|v^{(k-1,l)}\|_2})\|_2^2 \geq \|v^{(k-1,l)}\|_2^2 > A^2 > 0. \end{aligned} \tag{10}$$

In particular, by setting $\eta = k^{-0.5}$ and observing that $v^{(k,l)} = v^{(k-1,l)} - k^{-0.5}\nabla_{v^{(k-1,l)}} g(\frac{v^{(k-1,l)}}{\|v^{(k-1,l)}\|_2})$ we conclude that $\|v^{(k,l)}\|_2 > A$ so the step of induction is completed. □

Lemma 3. *Suppose conditions of Theorem 4 (and, hence, of Lemma 2) hold. Fix $k > 0$. According to (10) we can define a function $h : \mathbb{R} \to \mathbb{R}$ as*

$$h(\eta) = f\Big(H(v^{(1)}(\eta)) \dots H(v^{(L)}(\eta))\Big), \quad \forall l \in \{1, \dots, L\} :$$
$$v^{(l)}(\eta) = v^{(k-1,l)} - \eta \cdot \nabla_{v^{(k-1,l)}} \widetilde{f}\Big(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)})\Big).$$

Then

$$|\nabla h(\eta) - \nabla h(0)| \leq \mathcal{C}\widetilde{M}\eta,$$

where

$$\mathcal{C} = \frac{2L}{A^2}\Big(5\sqrt{6M_2(N+2)} + \sqrt{2M_2 + 48M_1^2(N+2)}(\sqrt{2(N+60)} + 8\sqrt{6N(N+2)})\Big),$$
$$\widetilde{M} = \sum_{l=1}^{L} \|\nabla_{v^{(k-1,l)}} \widetilde{f}(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}))\|_2^2.$$

Proof. Observe that due to (10) $\|v^{(l)}(\eta)\|_2 > A$. By applying a chain rule we deduce that

$$\nabla h(\eta) = -\sum_{l=1}^{L} \nabla_{v^{(k-1,l)}} \widetilde{f}(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}))^\top \nabla_{v^{(l)}} f(H(v^{(1)}(\eta)) \dots H(v^{(L)}(\eta))).$$

Next, we derive that

$$\begin{aligned}
|\nabla h(\eta) - \nabla h(0)| &= |\sum_{l=1}^{L} \nabla_{v^{(k-1,l)}} \widetilde{f}(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}))^\top \\
&\times \Big(\nabla_{v^{(l)}} f(H(v^{(1)}(\eta)) \dots H(v^{(L)}(\eta))) - \nabla_{v^{(k-1,l)}} f(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}))\Big)| \\
&\leq \sqrt{\widetilde{M}} \cdot \sqrt{\sum_{l=1}^{L} \|\nabla_{v^{(l)}} f(H(v^{(1)}(\eta)) \dots H(v^{(L)}(\eta))) - \nabla_{v^{(k-1,l)}} f(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}))\|_2^2} \qquad (11)
\end{aligned}$$

where we use the Cauchy-Schwarz inequality. Fix $l \in \{1, \dots, L\}$ and let $g'(v^{(l)})$, $g''(v^{(k-1,l)})$ be $f(H(v^{(1)}(\eta)) \dots H(v^{(L)}(\eta)))$ and $f(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}))$ represented as functions of $v^{(l)}$ and $v^{(k-1,l)}$ respectively. Then

$$\nabla_{v^{(l)}} f(H(v^{(1)}(\eta)) \dots H(v^{(L)}(\eta))) = \nabla_{v^{(l)}} g'(v^{(l)}) = \frac{1}{\|v^{(l)}\|_2}(I - s(v^{(l)})s(v^{(l)})^\top)\nabla_s g'(s(v^{(l)})),$$
$$\nabla_{v^{(k-1,l)}} f(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)})) = \nabla_{v^{(k-1,l)}} g''(v^{(k-1,l)})$$
$$= \frac{1}{\|v^{(k-1,l)}\|_2}(I - s(v^{(k-1,l)})s(v^{(k-1,l)})^\top)\nabla_s g''(s(v^{(k-1,l)})).$$

While l is fixed denote $v' = v^{(l)}$ and $v'' = v^{(k-1,l)}$. Then we have:

$$\begin{aligned}
&\|\nabla_{v^{(l)}} g'(v^{(l)}) - \nabla_{v^{(k-1,l)}} g''(v^{(k-1,l)})\|_2 = \|\frac{1}{\|v'\|_2}(I - s(v')s(v')^\top)\nabla_s g'(s(v')) \\
&- \frac{1}{\|v''\|_2}(I - s(v'')s(v'')^\top)\nabla_s g''(s(v''))\|_2 \\
&= \|\frac{1}{\|v'\|_2}(I - s(v')s(v')^\top)\nabla_s g'(s(v')) - \frac{1}{\|v''\|_2}(I - s(v')s(v')^\top)\nabla_s g'(s(v')) \\
&+ \frac{1}{\|v''\|_2}(I - s(v')s(v')^\top)\nabla_s g'(s(v')) - \frac{1}{\|v''\|_2}(I - s(v'')s(v'')^\top)\nabla_s g''(s(v''))\|_2 \\
&\leq \|\frac{1}{\|v'\|_2}(I - s(v')s(v')^\top)\nabla_s g'(s(v')) - \frac{1}{\|v''\|_2}(I - s(v')s(v')^\top)\nabla_s g'(s(v'))\|_2
\end{aligned}$$

$$
\begin{aligned}
&+ \|\frac{1}{\|v''\|_2}(I - s(v')s(v')^\top)\nabla_s g'(s(v')) - \frac{1}{\|v''\|_2}(I - s(v'')s(v'')^\top)\nabla_s g''(s(v''))\|_2 \\
&\leq |\frac{1}{\|v'\|_2} - \frac{1}{\|v''\|_2}|\|(I - s(v')s(v')^\top)\nabla_s g'(s(v'))\|_2 \\
&+ \frac{1}{\|v''\|_2}\|(I - s(v')s(v')^\top)\nabla_s g'(s(v')) - (I - s(v'')s(v'')^\top)\nabla_s g''(s(v''))\|_2 \\
&\leq \frac{1}{A^2}\|v' - v''\|_2\|\nabla_s g'(s(v'))\|_2 + \frac{1}{A}\|(I - s(v')s(v')^\top)\nabla_s g'(s(v')) - (I - s(v'')s(v'')^\top)\nabla_s g''(s(v''))\|_2 \\
&= \frac{1}{A^2}\|v' - v''\|_2\|\nabla_s g'(s(v'))\|_2 + \frac{1}{A}\|(I - s(v')s(v')^\top)\nabla_s g'(s(v')) \\
&- (I - s(v')s(v')^\top)\nabla_s g''(s(v'')) + (I - s(v')s(v')^\top)\nabla_s g''(s(v'')) - (I - s(v'')s(v'')^\top)\nabla_s g''(s(v''))\|_2 \\
&\leq \frac{1}{A^2}\|v' - v''\|_2\|\nabla_s g'(s(v'))\|_2 + \frac{1}{A}\|(I - s(v')s(v')^\top)\Big(\nabla_s g'(s(v')) - \nabla_s g''(s(v''))\Big)\|_2 \\
&+ \frac{1}{A}\|\Big((I - s(v')s(v')^\top) - (I - s(v'')s(v'')^\top)\Big)\nabla_s g''(s(v''))\|_2 \\
&\leq \frac{1}{A^2}\|v' - v''\|_2\|\nabla_s g'(s(v'))\|_2 + \frac{1}{A}\|\nabla_s g'(s(v')) - \nabla_s g''(s(v''))\|_2 \\
&+ \frac{1}{A}\|(s(v')s(v')^\top - s(v'')s(v'')^\top)\nabla_s g''(s(v''))\|_2 \\
&\leq \frac{1}{A^2}\|v' - v''\|_2\|\nabla_s g'(s(v'))\|_2 + \frac{1}{A}\|\nabla_s g'(s(v')) - \nabla_s g''(s(v''))\|_2 \\
&+ \frac{1}{A}\|s(v')s(v')^\top - s(v'')s(v'')^\top\|_2\|\nabla_s g''(s(v''))\|_2 \\
&\leq \frac{1}{A^2}\|v' - v''\|_2\|\nabla_s g'(s(v'))\|_2 + \frac{1}{A}\|\nabla_s g'(s(v')) - \nabla_s g''(s(v''))\|_2 \\
&+ \frac{1}{A}\|s(v')s(v')^\top - s(v')s(v'')^\top + s(v')s(v'')^\top - s(v'')s(v'')^\top\|_2\|\nabla_s g''(s(v''))\|_2 \\
&\leq \frac{1}{A^2}\|v' - v''\|_2\|\nabla_s g'(s(v'))\|_2 + \frac{1}{A}\|\nabla_s g'(s(v')) - \nabla_s g''(s(v''))\|_2 \\
&+ \frac{1}{A}\|s(v')(s(v') - s(v''))^\top\|_2\|\nabla_s g''(s(v''))\|_2 + \frac{1}{A}\|(s(v') - s(v''))s(v'')^\top\|_2\|\nabla_s g''(s(v''))\|_2 \\
&\leq \frac{1}{A^2}\|v' - v''\|_2\|\nabla_s g'(s(v'))\|_2 + \frac{1}{A}\|\nabla_s g'(s(v')) - \nabla_s g''(s(v''))\|_2 \\
&+ \frac{1}{A}\|s(v')\|_2\|s(v') - s(v'')\|_2\|\nabla_s g''(s(v''))\|_2 + \frac{1}{A}\|(s(v') - s(v'')\|_2\|s(v'')\|_2\|\nabla_s g''(s(v''))\|_2 \\
&\leq \frac{1}{A^2}\|v' - v''\|_2\|\nabla_s g'(s(v'))\|_2 + \frac{1}{A}\|\nabla_s g'(s(v')) - \nabla_s g''(s(v''))\|_2 + \frac{2}{A}\|s(v') - s(v'')\|_2\|\nabla_s g''(s(v''))\|_2 \\
&\leq \frac{1}{A^2}\|v' - v''\|_2(\|\nabla_s g'(s(v'))\|_2 + 4\|\nabla_s g''(s(v''))\|_2) + \frac{1}{A}\|\nabla_s g'(s(v')) - \nabla_s g''(s(v''))\|_2, \qquad (12)
\end{aligned}
$$

where we use submultiplicativity of the matrix norm $\|\cdot\|_2$ and that for any $v, v', v'' \in \mathcal{S}, x \in \mathbb{R}^N$: a) $I - s(v)s(v)^\top$ is an orthogonal projection matrix and, therefore, $\|I - s(v)s(v)^\top\|_2 \leq 1$, b) $|\frac{1}{\|v'\|_2} - \frac{1}{\|v''\|_2}| = \frac{1}{\|v'\|_2\|v''\|_2}|\|v'\|_2 - \|v''\|_2| \leq \frac{1}{A^2}\|v' - v''\|_2$ and c)

$$
\begin{aligned}
\|s(v') - s(v'')\|_2 = \|\frac{v'}{\|v'\|_2} - \frac{v''}{\|v''\|_2}\|_2 = \|\frac{v'}{\|v'\|_2} - \frac{v'}{\|v''\|_2} + \frac{v'}{\|v''\|_2} - \frac{v''}{\|v''\|_2}\|_2 \\
\leq |\frac{1}{\|v'\|_2} - \frac{1}{\|v''\|_2}|\|v'\|_2 + \frac{1}{\|v''\|_2}\|v' - v''\|_2 = \frac{1}{\|v''\|_2}|\|v'\|_2 - \|v''\|_2| + \frac{1}{\|v''\|_2}\|v' - v''\|_2 \\
\leq \frac{2}{\|v''\|_2}\|v' - v''\|_2 \leq \frac{2}{A}\|v' - v''\|_2.
\end{aligned}
$$

For $s \in \mathbb{R}^N, \|s\|_2 = 1$ let s_i denote i'th position of vector s, H_{j_1,j_2} denote (j_1, j_2)'th position of matrix H and $[\cdot]$ denote indicator. Then

$$
\nabla_{s_i} H(s)_{j_1,j_2} = \nabla_{s_i}(1 - 2\frac{s_{j_1}s_{j_2}}{\|s\|_2^2})
$$

$$= -2\frac{((s_{j_1}+s_{j_2})[j_1=i][j_2=i]+s_{j_1}[j_2=i][j_1\neq i]+s_{j_2}[j_1=i][j_2\neq i])\|s\|_2^2-2s_{j_1}s_{j_2}s_i}{\|s\|_2^4}$$

$$= 4s_{j_1}s_{j_2}s_i - 2(s_{j_1}[j_2=i]+s_{j_2}[j_1=i]).$$

We further obtain that

$$\begin{aligned}
\|\nabla_s H(s)\|_F^2 &= \sum_{1\le i,j_1,j_2\le N}(4s_{j_1}s_{j_2}s_i-2(s_{j_1}[j_2=i]+s_{j_2}[j_1=i]))^2 \\
&\le 3\sum_{1\le i,j_1,j_2\le N}(16s_{j_1}^2s_{j_2}^2s_i^2+4s_{j_1}^2[j_2=i]+4s_{j_2}[j_1=i]) \\
&= 48\sum_{j_1=1}^N s_{j_1}^2\sum_{j_2=1}^N s_{j_2}^2\sum_{i=1}^N s_i^2+12\sum_{1\le i,j_1\le N}s_{j_1}^2+12\sum_{1\le i,j_2\le N}s_{j_2}^2=24(N+2),
\end{aligned} \tag{13}$$

$$\begin{aligned}
&\|\nabla_{s'}H(s')-\nabla_{s''}H(s'')\|_F^2 = \sum_{1\le i,j_1,j_2\le N}(4s'_{j_1}s'_{j_2}s'_i-4s''_{j_1}s''_{j_2}s''_i-2(s'_{j_1}-s''_{j_1})[j_2=i] \\
&-2(s'_{j_2}-s''_{j_2})[j_1=i])^2 \le 3\sum_{1\le i,j_1,j_2\le N}(16(s'_{j_1}s'_{j_2}s'_i-s''_{j_1}s''_{j_2}s''_i)^2+4(s'_{j_1}-s''_{j_1})^2[j_2=i] \\
&+4(s'_{j_2}-s''_{j_2})^2[j_1=i]) \le 48\sum_{1\le i,j_1,j_2\le N}(s'_{j_1}s'_{j_2}s'_i-s'_{j_1}s''_{j_2}s''_i+s'_{j_1}s''_{j_2}s''_i-s''_{j_1}s''_{j_2}s''_i)^2 \\
&+4\sum_{1\le i,j_1\le N}(s'_{j_1}-s''_{j_1})^2+4\sum_{1\le i,j_2\le N}(s'_{j_2}-s''_{j_2})^2 \\
&\le 96\sum_{1\le i,j_1,j_2\le N}((s'_{j_1}s'_{j_2}s'_i-s'_{j_1}s''_{j_2}s''_i)^2+(s'_{j_1}s''_{j_2}s''_i-s''_{j_1}s''_{j_2}s''_i)^2)+8N\|s'-s''\|_2^2 \\
&= 96\sum_{1\le i,j_1,j_2\le N}(s'^2_{j_1}(s'_{j_2}s'_i-s''_{j_2}s''_i)^2+s''^2_{j_2}s''^2_i(s'_{j_1}-s''_{j_1})^2)+8N\|s'-s''\|_2^2 \\
&= 96\sum_{1\le i,j_1,j_2\le N}(s'^2_{j_1}(s'_{j_2}s'_i-s'_{j_2}s''_i+s'_{j_2}s''_i-s''_{j_2}s''_i)^2+s''^2_{j_2}s''^2_i(s'_{j_1}-s''_{j_1})^2)+8N\|s'-s''\|_2^2 \\
&= 96\sum_{1\le i,j_1,j_2\le N}(2s'^2_{j_1}s'^2_{j_2}(s'_i-s''_i)^2+2s'^2_{j_1}s''^2_i(s'_{j_2}-s''_{j_2})^2+s''^2_{j_2}s''^2_i(s'_{j_1}-s''_{j_1})^2) \\
&+8N\|s'-s''\|_2^2 \le 96\cdot 5\sum_{j_1=1}^N s'^2_{j_1}\sum_{j_2=1}^N s'^2_{j_2}\sum_{i=1}^N(s'_i-s''_i)^2+8N\|s'-s''\|_2 \le 8(60+N)\|s'-s''\|_2^2
\end{aligned} \tag{14}$$

where we use the Cauchy-Schwarz inequality and, in particular, that $(a+b)^2 \le 2(a^2+b^2)$ and $(a+b+c)^2 \le 3(a^2+b^2+c^2)$.

By Jensen's inequality, for every $X \in \mathcal{O}(N)$ we have $\|\nabla f(X)\|_F^2 = \|\mathbb{E}\nabla\widetilde{f}(X)\|_F^2 \le \mathbb{E}\|\nabla\widetilde{f}(X)\|_F^2 \le M_2^2$. By $X'(s)$, $X''(s)$ denote $H(v^{(1)}(\eta))\dots H(s(v'))\dots H(v^{(L)}(\eta))$ and $H(v^{(k-1,1)})\dots H(s(v''))\dots H(v^{(k-1,L)})$ as functions of $s(v')$ and $s(v'')$ respectively. Then

$$\begin{aligned}
\|\nabla_s X'(s)\|_F^2 &= \sum_{i=1}^N\|\nabla_{s_i}X'(s)\|_F^2=\sum_{i=1}^N\|H(v^{(1)}(\eta))\dots\nabla_{s_i}H(s(v'))\dots H(v^{(L)}(\eta))\|_F^2 \\
&= \sum_{i=1}^N\|\nabla_{s_i}H(s(v'))\|_F^2=\|\nabla_s H(s(v'))\|_F^2\le 24(N+2),
\end{aligned}$$

$$\begin{aligned}
&\|\nabla_s X'(s(v'))-\nabla_s X''(s(v''))\|_F = \|\nabla_s X'(s(v'))-\nabla_s X'(s(v''))+\nabla_s X'(s(v'')) \\
&-\nabla_s X''(s(v''))\|_F \le \|\nabla_s X'(s(v'))-\nabla_s X'(s(v''))\|_F+\|\nabla_s X'(s(v''))-\nabla_s X''(s(v''))\|_F \\
&= \sqrt{\sum_{i=1}^N\|H(v^{(1)}(\eta))\dots(\nabla_{s_i}H(s(v'))-\nabla_{s_i}H(s(v'')))\dots H(v^{(L)}(\eta))\|_F^2}+\|\nabla_s X'(s(v'')) \\
&-\nabla_s X''(s(v''))\|_F = \|\nabla_s H(s(v'))-\nabla_s H(s(v''))\|_F+\|\nabla_s X'(s(v''))-\nabla_s X''(s(v''))\|_F
\end{aligned}$$

$$\begin{aligned}&\le 2\sqrt{2(N+60)}\|s(v')-s(v'')\|_2 + \sum_{i=1}^{N}\|H(v^{(1)}(\eta))\dots\nabla_{s_i}H(s(v''))\dots H(v^{(L)}(\eta))\\ &\qquad -H(v^{(k-1,1)})\dots\nabla_{s_i}H(s(v''))\dots H(v^{(k-1,L)})\|_F\end{aligned}\tag{15}$$

For every $1 \le i \le N$ we have:

$$\|H(v^{(1)}(\eta))\dots\nabla_{s_i}H(s(v''))\dots H(v^{(L)}(\eta)) - H(v^{(k-1,1)})\dots\nabla_{s_i}H(s(v''))\dots H(v^{(k-1,L)})\|_F \tag{16}$$

$$\begin{aligned}&= \|H(v^{(1)}(\eta))\dots\nabla_{s_i}H(s(v''))\dots H(v^{(L)}(\eta)) - H(v^{(k-1,1)})H(v^{(2)}(\eta))\dots\nabla_{s_i}H(s(v''))\dots H(v^{(L)}(\eta))\\ &+H(v^{(k-1,1)})H(v^{(2)}(\eta))\dots\nabla_{s_i}H(s(v''))\dots H(v^{(L)}(\eta)) - H(v^{(k-1,1)})\dots\nabla_{s_i}H(s(v''))\dots H(v^{(k-1,L)})\|_F\\ &\le \|H(v^{(1)}(\eta))\dots\nabla_{s_i}H(s(v''))\dots H(v^{(L)}(\eta)) - H(v^{(k-1,1)})H(v^{(2)}(\eta))\dots\nabla_{s_i}H(s(v'))\dots H(v^{(L)}(\eta))\|_F\\ &+\|H(v^{(k-1,1)})\Big(H(v^{(2)}(\eta))\dots\nabla_{s_i}H(s(v''))\dots H(v^{(L)}(\eta)) - \dots\nabla_{s_i}H(s(v''))\dots H(v^{(k-1,L)})\Big)\|_F\\ &\le \|H(v^{(1)}(\eta)) - H(v^{(k-1,1)})\|_F\cdot\|H(v^{(2)}(\eta))\dots\nabla_{s_i}H(s(v''))\dots H(v^{(L)}(\eta))\|_F\\ &+\|H(v^{(2)}(\eta))\dots\nabla_{s_i}H(s(v''))\dots H(v^{(L)}(\eta)) - H(v^{(k-1,2)})\dots\nabla_{s_i}H(s(v''))\dots H(v^{(k-1,L)})\|_F\\ &\le \|H(v^{(1)}(\eta)) - H(v^{(k-1,1)})\|_F\|\nabla_{s_i}H(s(v''))\|_F\end{aligned}$$

$$+\|H(v^{(2)}(\eta))\dots\nabla_{s_i}H(s(v''))\dots H(v^{(L)}(\eta)) - H(v^{(k-1,2)})\dots\nabla_{s_i}H(s(v''))\dots H(v^{(k-1,L)})\|_F \tag{17}$$

$$\begin{aligned}&\le \dots \le \|\nabla_{s_i}H(s(v''))\|_F\sum_{l'=1}^{l-1}\|H(v^{(l')}(\eta)) - H(v^{(k-1,l')})\|_F\\ &+\|\nabla_{s_i}H(s(v''))\dots H(v^{(L)}(\eta)) - \nabla_{s_i}H(s(v''))\dots H(v^{(k-1,L)})\|_F,\end{aligned}$$

where $\dots$ correspond to repeating the reduction of type (16-17) to the term

$$\|H(v^{(2)}(\eta))\dots\nabla_{s_i}H(s(v''))\dots H(v^{(L)}(\eta)) - H(v^{(k-1,2)})\dots\nabla_{s_i}H(s(v''))\dots H(v^{(k-1,L)})\|_F$$

and so on until it becomes

$$\|\nabla_{s_i}H(s(v''))\dots H(v^{(L)}(\eta)) - \nabla_{s_i}H(s(v''))\dots H(v^{(k-1,L)})\|_F. \tag{18}$$

Then, one can repeat the reduction of type (16-17) to (18), but by extracting right-hand side reflections, so that (18) becomes $\|\nabla_{s_i}H(s(v'')) - \nabla_{s_i}H(s(v''))\|_F = 0$ and (16) is continued as

$$\begin{aligned}&\|H(v^{(1)}(\eta))\dots\nabla_{s_i}H(s(v'))\dots H(v^{(L)}(\eta)) - H(v^{(k-1,1)})\dots\nabla_{s_i}H(s(v''))\dots H(v^{(k-1,L)})\|_F\\ &\le \|\nabla_{s_i}H(s(v'))\|_F\sum_{l'\ne l}\|H(v^{(l')}(\eta)) - H(v^{(k-1,l')})\|_F + \|\nabla_{s_i}H(s(v')) - \nabla_{s_i}H(s(v''))\|_F\\ &\le \|\nabla_{s_i}H(s(v'))\|_F\sum_{l'=1}^{L}\|H(v^{(l')}(\eta)) - H(v^{(k-1,l')})\|_F\end{aligned}$$

We sum this inequality for $1 \le i \le N$, apply Cauchy-Schwarz inequality and use (13,14) to obtain that

$$\sum_{i=1}^{N}\|H(v^{(1)}(\eta))\dots\nabla_{s_i}H(s(v'))\dots H(v^{(L)}(\eta)) - H(v^{(k-1,1)})\dots\nabla_{s_i}H(s(v''))\dots H(v^{(k-1,L)})\|_F$$

$$\le \sqrt{N}\sqrt{\sum_{i=1}^{N}\|\nabla_{s_i}H(s(v'))\|_F^2}\sum_{l'=1}^{L}\|H(v^{(l')}(\eta)) - H(v^{(k-1,l')})\|_F \tag{19}$$

$$= 2\sqrt{N}\|\nabla_s H(s(v'))\|_F\sum_{l'=1}^{L}\|\frac{v^{(l')}(\eta)v^{(l')}(\eta)^\top}{\|v^{(l')}(\eta)\|_2^2} - \frac{v^{(k-1,l')}v^{(k-1,l')\top}}{\|v^{(k-1,l')}\|_2^2}\|_F \tag{20}$$

$$\le 4\sqrt{6N(N+2)}\sum_{l'=1}^{L}\|\frac{v^{(l')}(\eta)v^{(l')}(\eta)^\top}{\|v^{(l')}(\eta)\|_2^2} - \frac{v^{(k-1,l')}v^{(k-1,l')\top}}{\|v^{(k-1,l')}\|_2^2}\|_F \tag{21}$$

For each $1 \leq l' \leq L$ we have

$$\begin{gathered}
\|\frac{v^{(l')}(\eta)v^{(l')}(\eta)^\top}{\|v^{(l')}(\eta)\|_2^2} - \frac{v^{(k-1,l')}v^{(k-1,l')\top}}{\|v^{(k-1,l')}\|_2^2}\|_F \\
\leq \|\frac{v^{(l')}(\eta)v^{(l')}(\eta)^\top}{\|v^{(l')}(\eta)\|_2^2} - \frac{v^{(l')}(\eta)v^{(l')}(\eta)^\top}{\|v^{(l')}(\eta)\|_2\|v^{(k-1,l')}\|_2} + \frac{v^{(l')}(\eta)v^{(l')}(\eta)^\top}{\|v^{(l')}(\eta)\|_2\|v^{(k-1,l')}\|_2} - \frac{v^{(k-1,l')}v^{(k-1,l')\top}}{\|v^{(k-1,l')}\|_2^2}\|_F \\
\leq |\frac{1}{\|v^{(l')}(\eta)\|_2} - \frac{1}{\|v^{(k-1,l')}\|_2}| \cdot \frac{1}{\|v^{(l')}(\eta)\|_2}\|v^{(l')}(\eta)v^{(l')}(\eta)^\top\|_F \\
+\frac{1}{\|v^{(k-1,l')}\|_2}\|\frac{v^{(l')}(\eta)v^{(l')}(\eta)^\top}{\|v^{(l')}(\eta)\|_2} - \frac{v^{(k-1,l')}v^{(k-1,l')\top}}{\|v^{(k-1,l')}\|_2}\|_F \\
\leq |\|v^{(l')}(\eta)\|_2 - \|v^{(k-1,l')}\|_2| \\
\cdot\frac{1}{\|v^{(l')}(\eta)\|_2^2\|v^{(k-1,l')}\|_2}\|v^{(l')}(\eta)\|_2^2 + \frac{1}{\|v^{(k-1,l')}\|_2}\|\frac{v^{(l')}(\eta)v^{(l')}(\eta)^\top}{\|v^{(l')}(\eta)\|_2} - \frac{v^{(k-1,l')}v^{(k-1,l')\top}}{\|v^{(k-1,l')}\|_2}\|_F \\
\leq \frac{1}{A}\|v^{(l')}(\eta) - v^{(k-1,l')}\|_2 \\
+\frac{1}{\|v^{(k-1,l')}\|_2}\|\frac{v^{(l')}(\eta)v^{(l')}(\eta)^\top}{\|v^{(l')}(\eta)\|_2} - \frac{v^{(k-1,l')}v^{(k-1,l')\top}}{\|v^{(l')}(\eta)\|_2} + \frac{v^{(k-1,l')}v^{(k-1,l')\top}}{\|v^{(l')}(\eta)\|_2} - \frac{v^{(k-1,l')}v^{(k-1,l')\top}}{\|v^{(k-1,l')}\|_2}\|_F \\
\leq \frac{1}{A}\|v^{(l')}(\eta) - v^{(k-1,l')}\|_2 + \frac{1}{\|v^{(k-1,l')}\|_2\|v^{(l')}(\eta)\|_2}\|v^{(l')}(\eta)v^{(l')}(\eta)^\top - v^{(k-1,l')}v^{(k-1,l')\top}\|_F \\
+\frac{1}{\|v^{(k-1,l')}\|_2}|\frac{1}{\|v^{(l')}(\eta)\|_2} - \frac{1}{\|v^{(k-1,l')}\|_2}|\|v^{(k-1,l')}v^{(k-1,l')\top}\|_F \\
= \frac{1}{A}\|v^{(l')}(\eta) - v^{(k-1,l')}\|_2 + \frac{1}{\|v^{(k-1,l')}\|_2\|v^{(l')}(\eta)\|_2}\|v^{(l')}(\eta)v^{(l')}(\eta)^\top - v^{(k-1,l')}v^{(k-1,l')\top}\|_F \\
+\frac{1}{\|v^{(k-1,l')}\|_2^2\|v^{(l')}(\eta)\|_2}|\|v^{(l')}(\eta)\|_2 - \|v^{(k-1,l')}\|_2|\|v^{(k-1,l')}\|_2^2 \\
\leq \frac{1}{A}\|v^{(l')}(\eta) - v^{(k-1,l')}\|_2 \\
+\frac{1}{\|v^{(k-1,l')}\|_2\|v^{(l')}(\eta)\|_2}\|v^{(l')}(\eta)v^{(l')}(\eta)^\top - v^{(k-1,l')}v^{(k-1,l')\top}\|_F + \frac{1}{\|v^{(l')}(\eta)\|_2}\|v^{(l')}(\eta) - v^{(k-1,l')}\|_2 \\
\leq \frac{2}{A}\|v^{(l')}(\eta) - v^{(k-1,l')}\|_2 \\
+\frac{1}{\|v^{(k-1,l')}\|_2\|v^{(l')}(\eta)\|_2}\|v^{(l')}(\eta)v^{(l')}(\eta)^\top - v^{(l')}(\eta)v^{(k-1,l')\top} + v^{(l')}(\eta)v^{(k-1,l')\top} - v^{(k-1,l')}v^{(k-1,l')\top}\|_F \\
\leq \frac{2}{A}\|v^{(l')}(\eta) - v^{(k-1,l')}\|_2 \\
+\frac{1}{\|v^{(k-1,l')}\|_2\|v^{(l')}(\eta)\|_2}(\|v^{(l')}(\eta)(v^{(l')}(\eta) - v^{(k-1,l')})^\top\|_F + \|(v^{(l')}(\eta) - v^{(k-1,l')})v^{(k-1,l')\top}\|_F) \\
\leq \frac{2}{A}\|v^{(l')}(\eta) - v^{(k-1,l')}\|_2 + \frac{1}{\|v^{(k-1,l')}\|_2\|v^{(l')}(\eta)\|_2}(\|v^{(l')}(\eta)\|_2 + \|v^{(k-1,l')}\|_2)\|v^{(l')}(\eta) - v^{(k-1,l')}\|_2 \\
\leq \frac{2}{A}\|v^{(l')}(\eta) - v^{(k-1,l')}\|_2 + (\frac{1}{\|v^{(l')}(\eta)\|_2} + \frac{1}{\|v^{(k-1,l')}\|_2})\|v^{(l')}(\eta) - v^{(k-1,l')}\|_2 \leq \frac{4}{A}\|v^{(l')}(\eta) - v^{(k-1,l')}\|_2
\end{gathered}$$

We combine this with (15, 21) and conclude that

$$\begin{gathered}
\|\nabla_s X'(s(v')) - \nabla_s X''(s(v''))\|_F \leq 2\sqrt{2(N+60)}\|s(v') - s(v'')\|_2 \\
+4\sqrt{6N(N+2)}\sum_{l'=1}^{L}\left(\frac{4}{A}\|v^{(l')}(\eta) - v^{(k-1,l')}\|_2\right) \\
\leq \frac{2}{A}\sqrt{2(N+60)}\|v^{(l)} - v^{(k-1,l)}\|_2 + 4\sqrt{6N(N+2)}\sum_{l'=1}^{L}\frac{4}{A}\|v^{(l')}(\eta) - v^{(k-1,l')}\|_2
\end{gathered}$$

$$\leq \frac{2}{A}\left(\sqrt{2(N+60)}+8\sqrt{6N(N+2)}\right)\sum_{l'=1}^{L}\|v^{(l')}(\eta)-v^{(k-1,l')}\|_2$$

Next, we deduce that

$$\|\nabla_s g'(s(v'))\|_2^2=\sum_{i=1}^{N}(\nabla_{s_i} g'(s(v')))^2=\sum_{i=1}^{N}(\nabla_{s_i} f(X'(s(v'))))^2$$

$$=\sum_{i=1}^{N}\mathrm{Trace}(\nabla f(X'(s(v')))^\top\nabla_{s_i}X'(s(v')))^2\leq\sum_{i=1}^{N}\|\nabla f(X'(s(v')))\|_F^2\|\nabla_{s_i}X'(s(v'))\|_F^2$$

$$\leq M_2\sum_{i=1}^{N}\|\nabla_{s_i}X'(s(v'))\|_F^2=M_2\|\nabla_s X'(s(v'))\|_F^2\leq 24M_2(N+2).$$

Analogously it is derived that $\|\nabla_s g''(s(v''))\|_2^2\leq 24M_2(N+2)$. We proceed by observing that

$$\|\nabla_s g'(s(v'))-\nabla_s g''(s(v''))\|_2^2=\sum_{i=1}^{N}(\nabla_{s_i}g'(s(v'))-\nabla_{s_i}g''(s(v'')))^2$$

$$=\sum_{i=1}^{N}(\mathrm{Trace}(\nabla f(X'(s(v')))^\top\nabla_{s_i}X'(s(v')))-\mathrm{Trace}(\nabla f(X''(s(v'')))^\top\nabla_{s_i}X''(s(v''))))^2$$

$$=\sum_{i=1}^{N}(\mathrm{Trace}(\nabla f(X'(s(v')))^\top\nabla_{s_i}X'(s(v')))-\mathrm{Trace}(\nabla f(X'(s(v')))^\top\nabla_{s_i}X''(s(v'')))$$

$$+\mathrm{Trace}(\nabla f(X'(s(v')))^\top\nabla_{s_i}X''(s(v'')))-\mathrm{Trace}(\nabla f(X''(s(v'')))^\top\nabla_{s_i}X''(s(v''))))^2$$

$$\leq 2\sum_{i=1}^{N}(\mathrm{Trace}(\nabla f(X'(s(v')))^\top(\nabla_{s_i}X'(s(v'))-\nabla_{s_i}X''(s(v''))))^2$$

$$+\mathrm{Trace}((\nabla f(X'(s(v')))-\nabla f(X''(s(v''))))^\top\nabla_{s_i}X''(s(v'')))^2)$$

$$\leq 2\sum_{i=1}^{N}(\|\nabla f(X'(s(v')))\|_F^2\|\nabla_{s_i}X'(s(v'))-\nabla_{s_i}X''(s(v''))\|_F^2$$

$$+\|\nabla f(X'(s(v')))-\nabla f(X''(s(v'')))\|_F^2\|\nabla_{s_i}X''(s(v''))\|_F^2)$$

$$\leq 2M_2\sum_{i=1}^{N}\|\nabla_{s_i}X'(s(v'))-\nabla_{s_i}X''(s(v''))\|_F^2+2M_1^2\|X'(s(v'))-X''(s(v''))\|_F^2\sum_{i=1}^{N}\|\nabla_{s_i}X''(s(v''))\|_F^2$$

$$\leq(2M_2+2M_1^2\|\nabla_s X''(s(v''))\|_F^2)\|\nabla_s X'(s(v'))-\nabla_s X''(s(v''))\|_F^2$$

$$\leq(2M_2+48M_1^2(N+2))(\frac{2}{A}(\sqrt{2(N+60)}+8\sqrt{6N(N+2)})\sum_{l'=1}^{L}\|v^{(l')}(\eta)-v^{(k-1,l')}\|_2)^2.$$

We continue (12) and deduce that

$$\|\nabla_{v^{(l)}}g'(v^{(l)})-\nabla_{v^{(k-1,l)}}g''(v^{(k-1,l)})\|_2\leq\frac{10}{A^2}\sqrt{6M_2(N+2)}\|v'-v''\|_2$$

$$+\frac{2}{A^2}\sqrt{2M_2+48M_1^2(N+2)}(\sqrt{2(N+60)}+8\sqrt{6N(N+2)})\sum_{l'=1}^{L}\|v^{(l')}(\eta)-v^{(k-1,l')}\|_2$$

$$\leq\frac{\mathcal{C}}{L}\sum_{l'=1}^{L}\|v^{(l')}(\eta)-v^{(k-1,l')}\|_2\leq\frac{\mathcal{C}}{\sqrt{L}}\sqrt{\sum_{l'=1}^{L}\|v^{(l')}(\eta)-v^{(k-1,l')}\|_2^2}.$$

We plug the last inequality into (11) to obtain that

$$|\nabla h(\eta)-\nabla h(0)|\leq\frac{\mathcal{C}}{\sqrt{L}}\cdot\sqrt{\widetilde{M}}\cdot\sqrt{L\sum_{l'=1}^{L}\|v^{(l')}(\eta)-v^{(k-1,l')}\|_2^2}$$

$$= \mathcal{C}\sqrt{\widetilde{M}} \cdot \sqrt{\sum_{l=1}^{L} \| -\eta \nabla_{v^{(k-1,l)}} \widetilde{f}(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}))\|_2^2}$$

$$= \eta \cdot \mathcal{C}\sqrt{\widetilde{M}} \sqrt{\sum_{l=1}^{L} \|\nabla_{v^{(k-1,l)}} \widetilde{f}(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}))\|_2^2} = \mathcal{C}\widetilde{M}\eta.$$

□

Lemma 4. *Suppose conditions of Theorem 4 (and, consequently, of Lemma 2) hold. For any* $v^{(1)}, \dots, v^{(L)} \in \mathcal{S}$

$$\mathbb{E}[\sum_{l=1}^{L} \|\nabla_{v^{(l)}} \widetilde{f}(H(v^{(1)}) \dots H(v^{(L)}))\|_2^2] \leq \mathcal{D}, \quad \mathcal{D} = \frac{24}{A^2} N(N+2)LM_2.$$

Proof. For each $1 \leq l \leq L$ we have

$$\|\nabla_{v^{(l)}} \widetilde{f}(H(v^{(1)}) \dots H(v^{(L)}))\|_2^2 = \sum_{i=1}^{N} (\nabla_{v_i^{(l)}} \widetilde{f}(H(v^{(1)}) \dots H(v^{(L)})))^2$$

$$= \sum_{i=1}^{N} \mathrm{Trace}(\nabla \widetilde{f}(H(v^{(1)}) \dots H(v^{(L)}))^\top \nabla_{v_i^{(l)}} \Big(H(v^{(1)}) \dots H(v^{(L)}) \Big))^2$$

$$\leq \sum_{i=1}^{N} \|\nabla \widetilde{f}(H(v^{(1)}) \dots H(v^{(L)}))\|_F^2 \cdot \|H(v^{(1)}) \dots \nabla_{v_i^{(l)}} H(v^{(l)}) \dots H(v^{(L)})\|_F^2$$

$$= \|\nabla \widetilde{f}(H(v^{(1)}) \dots H(v^{(L)}))\|_F^2 \sum_{i=1}^{N} \|\nabla_{v_i^{(l)}} H(v^{(l)})\|_F^2 = \|\nabla \widetilde{f}(H(v^{(1)}) \dots H(v^{(L)}))\|_F^2 \|\nabla_{v^{(l)}} H(v^{(l)})\|_F^2$$

$$= \|\nabla \widetilde{f}(H(v^{(1)}) \dots H(v^{(L)}))\|_F^2 \sum_{i=1}^{N} \|\nabla_{v_i^{(l)}} H(s(v^{(l)}))\|_F^2$$

$$= \|\nabla \widetilde{f}(H(v^{(1)}) \dots H(v^{(L)}))\|_F^2 \sum_{i=1}^{N} \| \sum_{j=1}^{N} \nabla_{v_i^{(l)}} s_j(v^{(l)}) \nabla_{s_j} H(s(v^{(l)}))\|_F^2$$

$$\leq \|\nabla \widetilde{f}(H(v^{(1)}) \dots H(v^{(L)}))\|_F^2 \sum_{i=1}^{N} \|\nabla_{v_i^{(l)}} s(v^{(l)})\|_2^2 \|\nabla_s H(s(v^{(l)}))\|_F^2$$

$$= \|\nabla \widetilde{f}(H(v^{(1)}) \dots H(v^{(L)}))\|_F^2 \|\nabla_{v^{(l)}} s(v^{(l)})\|_F^2 \|\nabla_s H(s(v^{(l)}))\|_F^2$$

$$\leq \|\nabla \widetilde{f}(H(v^{(1)}) \dots H(v^{(L)}))\|_F^2 \frac{1}{\|v^{(l)}\|_2^2} \|I - s(v^{(l)}) s(v^{(l)})^\top\|_F^2 \cdot 24(N+2)$$

$$\leq \frac{24}{A^2} N(N+2) \|\nabla \widetilde{f}(H(v^{(1)}) \dots H(v^{(L)}))\|_F^2$$

where we use $\|I - s(v^{(l)}) s(v^{(l)})^\top\|_F^2 \leq N$ because $I - s(v^{(l)}) s(v^{(l)})^\top$ is an orthogonal projection matrix. Next, we obtain that

$$\mathbb{E} \sum_{l=1}^{L} \|\nabla_{v^{(k-1,l)}} \widetilde{f}(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}))\|_2^2 \leq \frac{24}{A^2} N(N+2) \cdot \mathbb{E} L \|\nabla \widetilde{f}(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}))\|_F^2 \leq \mathcal{D}.$$

□

Theorem 4 proof. As shown by Lemma 2, all step sizes are well-defined. We adapt the proof of Theorem 4.10 from (Bottou et al., 2016). We consider a step k and deduce from Lemma 3 that

$$h(k^{-0.5}) - h(0) - k^{-0.5} \nabla h(0) = \int_0^{k^{-0.5}} (\nabla h(\eta) - \nabla h(0)) d\eta \leq \int_0^{k^{-0.5}} |\nabla h(\eta) - \nabla h(0)| d\eta$$

$$\leq \mathcal{C}\widetilde{M} \int_0^{k^{-0.5}} \eta d\eta = \frac{\mathcal{C}\widetilde{M}k^{-1}}{2}.$$

By expanding h's definition, we deduce

$$f\Big(H(v^{(k,1)}) \dots H(v^{(k,L)}) \Big) - f\Big(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}) \Big) \leq \frac{\mathcal{C}\widetilde{M}k^{-1}}{2}$$
$$-k^{-0.5} \sum_{l=1}^{L} \nabla_{v^{(k-1,l)}} \widetilde{f}(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}))^\top \nabla_{v^{(k-l,l)}} f(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)})).$$

Take expectation conditioned on $\mathcal{F}_k$ – a σ-algebra associated with $\{\{v^{(k',1)}, \dots, v^{(k',L)}\}\}_{k'=1}^{k-1}$:

$$\mathbb{E}[f\Big(H(v^{(k,1)}) \dots H(v^{(k,L)}) \Big) | \mathcal{F}_k] - f\Big(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}) \Big) \leq \frac{\mathcal{C}\mathbb{E}[\widetilde{M}|\mathcal{F}_k]k^{-1}}{2}$$
$$-k^{-0.5} \sum_{l=1}^{L} \mathbb{E}[\nabla_{v^{(k-1,l)}} \widetilde{f}(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)})) | \mathcal{F}_k]^\top \times \nabla_{v^{(k-1,l)}} f(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)})).$$

By $\widetilde{f}$'s definition we have

$$\mathbb{E}[\nabla_{v^{(k-1,l)}} \widetilde{f}(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)})) | \mathcal{F}_k] = \nabla_{v^{(k-1,l)}} f(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}))$$

and, therefore,

$$\mathbb{E}[f\Big(H(v^{(k,1)}) \dots H(v^{(k,L)}) \Big) | \mathcal{F}_k] - f\Big(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}) \Big) \leq \frac{\mathcal{C}\mathbb{E}[\widetilde{M}|\mathcal{F}_k]k^{-1}}{2}$$
$$-k^{-0.5} \sum_{l=1}^{L} \|\nabla_{v^{(k-1,l)}} f(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}))\|_2^2. \tag{22}$$

Next, we combine (22) and Lemma 4, applied to $\widetilde{M}$, to obtain that

$$\mathbb{E}[f\Big(H(v^{(k,1)}) \dots H(v^{(k,L)}) \Big) | \mathcal{F}_k] - f\Big(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}) \Big) \leq \frac{\mathcal{C}\mathcal{D}k^{-1}}{2}$$
$$-k^{-0.5} \sum_{l=1}^{L} \|\nabla_{v^{(k-1,l)}} f(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}))\|_2^2.$$

Take full expectation and regroup:

$$k^{-0.5} \mathbb{E} \sum_{l=1}^{L} \|\nabla_{v^{(k-1,l)}} f(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}))\|_2^2 \leq \mathbb{E} f\Big(H(v^{(k-1,1)}) \dots H(v^{(k-1,L)}) \Big)$$
$$-\mathbb{E} f\Big(H(v^{(k,1)}) \dots H(v^{(k,L)}) \Big) + \frac{\mathcal{C}\mathcal{D}k^{-1}}{2}.$$

For $K > 0$ take a sum for $1 \leq k \leq K$:

$$\sum_{k'=1}^{K} k'^{-0.5} \mathbb{E} \sum_{l=1}^{L} \|\nabla_{v^{(k'-1,l)}} f(H(v^{(k'-1,1)}) \dots H(v^{(k'-1,L)}))\|_2^2 \leq f\Big(H(v^{(0,1)}) \dots H(v^{(0,L)}) \Big)$$
$$-\mathbb{E} f\Big(H(v^{(K,1)}) \dots H(v^{(K,L)}) \Big) + \sum_{k'=1}^{K} \frac{\mathcal{C}\mathcal{D}k'^{-1}}{2}.$$

f is continuous on a compact domain $\mathcal{O}(N)$, hence there exists a minimal value f^* of f on $\mathcal{O}(N)$. We continue and derive that

$$\sum_{k'=1}^{K} k'^{-0.5} \mathbb{E} \sum_{l=1}^{L} \|\nabla_{v^{(k'-1,l)}} f(H(v^{(k'-1,1)}) \dots H(v^{(k'-1,L)}))\|_2^2 \leq f\Big(H(v^{(0,1)}) \dots H(v^{(0,L)}) \Big) - f^* + \sum_{k'=1}^{K} \frac{\mathcal{C}\mathcal{D}k'^{-1}}{2},$$

$$
\begin{aligned}
&\min_{0 \leq k' < K} \mathbb{E} \sum_{l=1}^{L} \| \nabla_{v^{(k'-1,l)}} f(H(v^{(k'-1,1)}) \dots H(v^{(k'-1,L)})) \|_2^2 \\
&\leq \frac{1}{\sum_{k'=1}^{K} k'^{-0.5}} \sum_{k'=1}^{K} k'^{-0.5} \mathbb{E} \sum_{l=1}^{L} \| \nabla_{v^{(k'-1,l)}} f(H(v^{(k'-1,1)}) \dots H(v^{(k'-1,L)})) \|_2^2 \\
&\leq \frac{1}{\sum_{k'=1}^{K} k'^{-0.5}} (f \left(H(v^{(0,1)}) \dots H(v^{(0,L)}) \right) - f^*) + \frac{\mathcal{C}\mathcal{D}}{2} \frac{\sum_{k'=1}^{K} k'^{-1}}{\sum_{k'=1}^{K} k'^{-0.5}}.
\end{aligned}
$$

The proof is concluded by observing that $\sum_{k'=1}^{K} k'^{-0.5} = \Omega(K^{0.5})$ and $\sum_{k'=1}^{K} k'^{-1} = O(\log K) = o(K^{\epsilon})$ for any $\epsilon > 0$. □