A. Riemannian Geometry and Lie Groups

In this section we aim to give a short summary of the basics of classical Riemannian geometry and Lie group theory needed for our proofs in following sections. The standard reference for classical Riemannian geometry is do Carmo's book (do Carmo, 1992). An elementary introduction to Lie group theory with an emphasis on concrete examples from matrix theory can be found in (Hall, 2015).

A.1. Riemannian geometry

A Riemannian manifold is a smooth manifold $\mathcal{M}$ equipped with a smooth metric $\langle \cdot, \cdot \rangle_p \colon T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$ which is a positive definite inner product for every $p \in \mathcal{M}$. We will omit the dependency on the point p whenever it is clear from the context. We will work with finite dimensional manifolds. Given a metric, we can define the length of a curve γ on the manifold as $L(\gamma) := \int_a^b \sqrt{\langle \gamma'(t), \gamma'(t) \rangle}\, \mathrm{d}t$. The distance between two points is the infimum of the lengths of the piece-wise smooth curves on $\mathcal{M}$ connecting them. When the manifold is connected, this defines a distance function that turns the manifold into a metric space.

An *affine connection* ∇ on a smooth manifold is a bilinear application that maps two vector fields X, Y to a new one $\nabla_X Y$ such that it is linear in X, and linear and Leibnitz in Y.

Connections give a notion of variation of a vector field along another vector field. In particular, the covariant derivative is the restriction of the connection to a curve. In particular, we can define the notion of parallel transport of vectors. We say that a vector field Z is *parallel* along a curve γ if $\nabla_{\gamma'} Z = 0$ where $\gamma' := \mathrm{d}\gamma(\frac{\mathrm{d}}{\mathrm{d}t})$. Given initial conditions $(p, v) \in T\mathcal{M}$ there exists locally a unique parallel vector field Z along γ such that $Z(p) = v$. $Z(\gamma(t))$ is sometimes referred to as the *parallel transport of v along γ*.

We say that a connection is *compatible with the metric* if for any two parallel vector fields X, Y along γ, their scalar product is constant. In other words, the connection preserves the angle between parallel vector fields. We say that a connection is *torsion-free* if $\nabla_X Y - \nabla_Y X = [X, Y] := XY - YX$. In a Riemannian manifold, there exists a unique affine connection such that it is compatible with the metric and that is also torsion-free. We call this distinguished connection the *Levi-Civita connection*.

A geodesic is defined as a curve such that its tangent vectors are covariantly constant along itself, $\nabla_{\gamma'} \gamma' = 0$. It is not true in general that given two points in a manifold there exists a geodesic that connects them. However, the *Hopf-Rinow theorem* states that this is indeed the case if the manifold is connected and complete as a metric space. The manifolds that we will consider are all connected and complete.

At every point p in our manifold we can define the *Riemannian exponential map* $\exp_p \colon T_p\mathcal{M} \to \mathcal{M}$, which maps a vector v to $\gamma(1)$ where γ is the geodesic such that $\gamma(0) = p$, $\gamma'(0) = v$. In a complete manifold, another formulation of the *Hopf-Rinow theorem* says that the exponential map is defined on the whole tangent space for every $p \in \mathcal{M}$. We also have that the Riemannian exponential map maps diffeomorphically a neighborhood around zero on the tangent space to a neighborhood of the point on which it is defined.

A map between two Riemannian manifolds is called a (local) isometry if it is a (local) diffeomorphism and its differential respects the metric.

A.2. Lie groups

A Lie group is a smooth manifold equipped with smooth group multiplication and inverse. Examples of Lie groups are the Euclidean space equipped with its additive group structure and the general linear group GL of a finite dimensional vector space given by the invertible linear endomorphisms of the space equipped with the composition of morphisms. We say that a Lie group is a matrix Lie group if it is a closed subgroup of some finite-dimensional general linear group.

Lie groups act on themselves via the left translations given by $L_g(x) = gx$ for $g, x \in G$. A vector field X is called *left invariant* if $(\mathrm{d}L_g)(X) = X \circ L_g$. A left invariant vector field is uniquely determined by its value at the identity of the group. This identification gives us a way to identify the tangent space at a point of the group with the tangent space at the identity. We call the tangent space at the identity *the Lie algebra of G* and we denote it by $\mathfrak{g}$.

For every vector $v \in \mathfrak{g}$ there exists a unique curve $\gamma \colon \mathbb{R} \to G$ such that γ is the integral curve of the left-invariant vector field defined by v such that $\gamma(0) = e$. This curve is a Lie group homomorphism and we call it the *Lie exponential*. It is also

the integral curve of the right-invariant vector field with initial vector v.

We say that $c_g(x) = gxg^{-1}$ for $g, x \in G$ is an *inner automorphism of* G. Its differential at the identity is the *adjoint representation of* G, $\mathrm{Ad}\colon G \to \mathrm{GL}(\mathfrak{g})$ defined as $\mathrm{Ad}_g(X) := (\mathrm{d}c_g)_e(X)$ for $g \in G$, $X \in \mathfrak{g}$. The differential at the identity of Ad is called the *adjoint representation of* $\mathfrak{g}$, $\mathrm{ad}\colon \mathfrak{g} \to \mathrm{End}(\mathfrak{g})$ defined as $\mathrm{ad}_X(Y) := (\mathrm{d}\,\mathrm{Ad})_e(X)(Y)$. We say that $\mathrm{ad}_X(Y)$ is the *Lie bracket of* $\mathfrak{g}$ and we denote it by $[X, Y]$. For a matrix Lie group we have $\mathrm{Ad}_g(X) = gXg^{-1}$ and $\mathrm{ad}_X(Y) = XY - YX$.

A (complex) representation of a group is a continuous group homomorphism $\rho\colon G \to \mathrm{GL}(n, \mathbb{C})$. An injective representation is called *faithful*. The inclusion is a faithful representation for any matrix Lie group. On a compact Lie group, $\rho(g)$ is diagonalizable for every $g \in G$.

A Riemannian metric on G is said to be bi-invariant if it turns left and right translations into isometries. We have that every compact Lie group admits a bi-invariant metric. An example of a bi-invariant metric on the group of orthogonal matrices with positive determinant $\mathrm{SO}(n)$ is that inherited from $\mathbb{R}^{n\times n}$, namely the canonical metric $\langle X, Y\rangle = \mathrm{tr}(X^\intercal Y)$. The same happens in the unitary case, but changing the transpose for a conjugate transpose $\langle X, Y\rangle = \mathrm{tr}(X^* Y)$. Furthermore, every Lie group that admits a bi-invariant metric is a homogeneous Riemannian manifold—there exists an isometry between that takes any point to any other point—, and hence, complete.

B. Retractions

We take a deeper look into the concept of a retraction, which helps understanding the correctness of the approach used to optimize on $\mathrm{SO}(n)$ and $\mathrm{U}(n)$ presented in (Wisdom et al., 2016; Vorontsov et al., 2017).

The concept of a retraction is a relaxation of that of the Riemannian exponential map.

Definition B.1 (Retraction). A retraction is a map

$$\begin{aligned} r\colon T\mathcal{M} &\to \mathcal{M} \\ (x, v) &\mapsto r_x(v) \end{aligned}$$

such that

$$r_x(0) = x \qquad \text{and} \qquad (\mathrm{d}r_x)_0 = \mathrm{Id}$$

where Id is the identity map.

In other words, when $\mathcal{M}$ is a Riemannian manifold, r is a first order approximation of the Riemannian exponential map.

It is clear that the exponential map is a retraction. For manifolds embedded in the Euclidean space with the metric induced by that of the Euclidean space, the following proposition gives us a simple way to construct a rather useful family of retractions—those used in projected gradient descent.

Proposition B.2. *Let $\mathcal{M}$ be an embedded submanifold of $\mathbb{R}^n$, then for a differentiable surjective projection $\pi\colon \mathbb{R}^n \to \mathcal{M}$, that is, $\pi \circ \pi = \pi$, the map*

$$\begin{aligned} r\colon T\mathcal{M} &\to \mathcal{M} \\ (x, v) &\mapsto \pi(x + v) \end{aligned}$$

is a retraction, where we are implicitly identifying $T_x\mathcal{M} \subseteq T_x\mathbb{R}^n \cong \mathbb{R}^n$.

Proof. From π being a surjective projection we have that $\pi(x) = x$ for every $x \in \mathcal{M}$, which implies the first condition of the definition of retraction.

Another way of seeing the above is saying that $\pi|_{\mathcal{M}} = \mathrm{Id}$. This implies that, for every $x \in \mathcal{M}$, $(\mathrm{d}\pi)_x = \mathrm{Id}|_{T_x\mathcal{M}}$. By the chain rule, since the differential of $v \mapsto x + v$ is the identity as well we get the second condition. □

This proposition lets us see projected Riemannian gradient descent as an specific case of Riemannian gradient descent with a specific retraction. A corollary of this proposition that allows r to be defined just in those vectors of the form $x + v$ with $(x, v) = T\mathcal{M}$ lets us construct specific examples of retractions:

Example B.3 (Sphere). The function

$$r_x(v) = \frac{x+v}{\|x+v\|}$$

for $v \in T_x\mathbb{S}^n$ is a retraction.

Example B.4 (Special orthogonal group). Recall that for $A \in \mathrm{SO}(n)$,

$$T_A(\mathrm{SO}(n)) = \{X \in \mathbb{R}^{n\times n} \mid A^\intercal X + X^\intercal A = 0\},$$

then, for an element of $T\,\mathrm{SO}(n)$ we can define the map given by Proposition B.2. In this case, the projection $\pi(X)$ for a matrix with singular value decomposition $X = U\Sigma V^\intercal$ is $\pi(X) = UV^\intercal$.

This projection is nothing but the orthogonal projection from $\mathbb{R}^{n\times n}$ onto $\mathrm{SO}(n)$ when equipped with the canonical metric.

The two examples above are examples of orthogonal projections. The manifolds being considered are embedded into a Euclidean space and they inherit its metric. The projections here are orthogonal projections on the ambient space. On the other hand, Proposition B.2 does not require the projections to be orthogonal.

Different examples of retractions can be found in (Absil et al., 2009), Example $4.1.2$.

C. Comparing Riemannian gradient descent and the exponential parametrization

The proofs in this section are rather technical and general so that they apply to a wide variety of manifolds such as $\mathrm{SO}(n)$, $\mathrm{U}(n)$, or the symplectic group. Even though we do not explore applications of optimizing over other compact matrix Lie groups, we state the results in full generality. Having this in mind, we will first motivate this section with the concrete example that we study in the applications of this paper: $\mathrm{SO}(n)$.

Example C.1. Let $f\colon \mathbb{R}^{n\times n} \to \mathbb{R}$ be a function defined on the space of matrices. Our task is to solve the problem

$$\min_{B\in\mathrm{SO}(n)} f(B).$$

A simple way to approach this problem would be to apply Riemannian gradient descent to it. Let A be a skew-symmetric matrix an let $B = e^A \in \mathrm{SO}(n)$, where e^A denotes the exponential of matrices. We will suppose that we are working with the canonical metric on $\mathbb{R}^{n\times n}$, namely $\langle X, Y\rangle = \mathrm{tr}(X^\intercal Y)$. We will denote by $\nabla f(B)$ the gradient of the function f at the matrix B, and by $\mathrm{grad}\, f(B) \in T_B\,\mathrm{SO}(n)$ the gradient associated to the restriction $f|_{\mathrm{SO}(n)}$ with respect to the induced metric.

Riemannian gradient descent works by following the geodesic defined by the direction $-\mathrm{grad}\, f(B)$ at the point B. In the words of the Riemannian exponential map at B, if we have a learning rate $\eta > 0$, the update rule will be given by

$$B \leftarrow \exp_B(-\eta\,\mathrm{grad}\, f(B)).$$

The tangent space to $\mathrm{SO}(n)$ at a matrix B is

$$T_B\,\mathrm{SO}(n) = \{X \in \mathbb{R}^{n\times n} \mid B^\intercal X + X^\intercal B = 0\}$$

and it is easy to check that the orthogonal projection with respect to the canonical metric onto this space is given by

$$\begin{aligned}\pi_B\colon \mathbb{R}^{n\times n} &\to T_B\,\mathrm{SO}(n)\\ X &\mapsto \tfrac{1}{2}(X - BX^\intercal B)\end{aligned}$$

Since the gradient on the manifold is just the tangent component of the gradient on the ambient space, we have that

$$\mathrm{grad}\, f(B) = \tfrac{1}{2}(\nabla f(B) - B\nabla f(B)^\intercal B),$$

and since multiplying by an orthogonal matrix constitutes an isometry, in order to compute the Riemannian exponential map we can transport the vector from T_BG to $T_{\mathrm{I}}G$, compute the exponential at the identity using the exponential of matrices and then transport the result back. In other words,

$$\exp_B(X) = B\exp(B^\intercal X) \qquad \forall X \in T_B\,\mathrm{SO}(n).$$

Putting everything together, the Riemannian gradient descent update rule for $\mathrm{SO}(n)$ is given by

$$B \leftarrow e^{A} e^{-\eta B^\intercal \operatorname{grad} f(B)}.$$

The other update rule that we have is the one given by the exponential parametrization

$$B \leftarrow e^{A - \eta \nabla (f \circ \exp)(A)}.$$

This is nothing but the gradient descent update rule applied to the problem

$$\min_{A \in \mathrm{Skew}(n)} f(\exp(A)).$$

Both of these rules follow geodesic flows for two metrics whenever A is in a neighborhood of the identity on which $\exp$ is a diffeomorphism (*cf.*, Appendix D). A natural question that arises is whether these metrics are the same. A less restrictive question would be whether this second optimization procedure defines a retraction or whether their gradient flow is completely different.

In the sequel, we will see that for $\mathrm{SO}(n)$ these two optimization methods give raise to two rather different metrics. We will explicitly compute the quantity $\nabla(f \circ \exp)(A)$, and we will give necessary and sufficient conditions equivalent under which these two optimization methods agree.

C.1. Optimization on Lie Groups with Bi-invariant Metrics

In this section we expose the theoretical part of the paper. The first part of this section is classic and can be found, for example, in Milnor (Milnor, 1976). We present it here for completeness. The results Proposition C.11 and Theorem C.12 are novel.

Remark. Throughout this section the operator $(-)^*$ will have two different meanings. It can be either the pullback of a form along a function or the adjoint of a linear operator on a vector space with an inner product. Although the two can be distinguished in many situations, we will explicitly mention to which one we are referring whenever it may not be clear from the context. Note that when we are on a matrix Lie group equipped with the product $\langle X, Y \rangle = \operatorname{tr}(X^* Y)$, the adjoint of a linear operator is exactly its conjugate transpose, hence the notation.

When we deal with an abstract group we will denote the identity element as e. If the group is a matrix Lie group, we will sometimes refer to it as I.

We start by recalling the definition of our object of study.

Definition C.2 (Invariant metric on a Lie Group). A Riemannian metric on a Lie group G is said to be left (resp. right) invariant if it makes left (resp. right) translations into isometries. Explicitly, it is so if for every $g \in G$, and the metric α we have that $\alpha_g = L_{g^{-1}}^* \alpha_e$ (resp. $\alpha_g = R_{g^{-1}}^* \alpha_e$).

A bi-invariant metric is a metric that is both left and right-invariant.

We can construct a bi-invariant metric on a Lie group by using the *averaging trick*.

Proposition C.3 (Bi-invariant metric on compact Lie groups). *A compact Lie group G admits a bi-invariant metric.*

Proof. Let n be the dimension of G and let μ_e be a non-zero n-form at $\mathfrak{g}$. This form is unique up to a multiplicative constant. We can then extend it to the whole G by pulling it back along R_g defining $\mu_g := R_{g^{-1}}^* \mu_e$. This makes it into a right-invariant n-form on the manifold, which we call the *right Haar measure*.

Let $(\cdot, \cdot)$ be an inner product on $\mathfrak{g}$. We can turn this inner product into an Ad-invariant inner product on $\mathfrak{g}$ by averaging it over the elements of the group using the right Haar measure

$$\langle u, v \rangle = \int_G (\mathrm{Ad}_g(u), \mathrm{Ad}_g(v))\, \mu(\mathrm{d}g).$$

Note that this integral is well defined since G is compact. The Ad-invariance follows from the right-invariance of μ

$$\langle \mathrm{Ad}_h(u), \mathrm{Ad}_h(v) \rangle = \int_G (\mathrm{Ad}_{gh}(u), \mathrm{Ad}_{gh}(v))\, \mu(\mathrm{d}g) = \langle u, v \rangle.$$

Finally, we can extend this product to the whole group by pulling back the inner product along L_g, that is, if we denote the metric by α, $\alpha_g = L_{g^{-1}}^* \alpha_e$. This automatically makes it into a left-invariant metric. But since it is Ad-invariant at the identity, we have that for every $g, h \in G$

$$R_g^* \alpha_{hg} = R_g^* L_{g^{-1}h^{-1}}^* \alpha_e = \mathrm{Ad}_{g^{-1}}^* L_{h^{-1}}^* \alpha_e = L_{h^{-1}}^* \alpha_e = \alpha_h$$

and the metric is also right-invariant, finishing the proof. □

If the group is abelian, the construction above is still valid without the need of the averaging trick, since Ad is the identity map, so every inner product is automatically Ad-invariant.

It turns out that these examples and their products exhaust all the Lie groups that admit a bi-invariant metric. We include this result for completeness, even though we will not use it.

Theorem C.4 (Classification of groups with bi-invariant metrics). *A Lie group admits a bi-invariant metric if and only if it is isomorphic to $G \times H$ with G compact and H abelian.*

Proof. (Milnor, 1976) Lemma 7.5. □

Lie groups, when equipped with a bi-invariant metric are rather amenable from the Riemannian geometry perspective. This is because it is possible to reduce many computations on them to matrix algebra, rather than the usual systems of differential equations that one encounters when dealing with arbitrary Riemannian manifolds.

The following proposition will come in handy later.

Lemma C.5. *If an inner product on $\mathfrak{g}$ is Ad-invariant then*

$$\langle Y, \mathrm{ad}_X(Z) \rangle = -\langle \mathrm{ad}_X(Y), Z \rangle \qquad \forall X, Y, Z \in \mathfrak{g}.$$

In other words, the adjoint of the map ad_X with respect to the inner product is $-\mathrm{ad}_X$. We say that ad is skew-adjoint and we write $\mathrm{ad}_X^* = -\mathrm{ad}_X$.

Proof. We have that, by definition

$$\mathrm{ad}_X(Y) = \frac{\mathrm{d}}{\mathrm{d}t}(\mathrm{Ad}_{\exp(tX)}(Y)) \mid_0,$$

so, deriving the equation

$$\langle \mathrm{Ad}_{\exp(tX)}(Y), \mathrm{Ad}_{\exp(tX)}(Z) \rangle = \langle Y, Z \rangle$$

with respect to t we get the result. □

With this result in hand, we can prove a rather useful relation between the geometry of the Lie group and its algebraic structure.

Proposition C.6. *Let G be a Lie group equipped with a bi-invariant metric. If X, Y are left-invariant vector fields, we have that their Levi-Civita connection and their sectional curvature are given by*

$$\nabla_X Y = \frac{1}{2}[X, Y]$$
$$\kappa(X, Y) = \frac{1}{4}\|[X, Y]\|^2.$$

The sectional curvature formula holds whenever X and Y are orthonormal.

Proof. For left-invariant vector fields X, Y, Z, the Koszul formula gives

$$\langle \nabla_X Y, Z \rangle = \frac{1}{2}(X\langle Y, Z \rangle + Y\langle X, Z \rangle - Z\langle X, Y \rangle + \langle [X, Y], Z \rangle - \langle [X, Z], Y \rangle - \langle [Y, Z], X \rangle).$$

The three first terms on the right vanish, since the angle between invariant vector fields is constant. Reordering the last three terms, using Lemma C.5, the fact that the Lie bracket is antisymmetric, and since invariant vector fields form a basis of the Lie algebra, the formula for the connection follows. Now, the curvature tensor is given by

$$R(X,Y)Z = \nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z = \frac{1}{4}([X,[Y,Z]] - [Y,[X,Z]] - 2[[X,Y],Z]) = \frac{1}{4}[Z,[X,Y]].$$

So the sectional curvature for X, Y orthonormal is given by

$$\kappa(X,Y) = \langle R(X,Y)Y, X\rangle = \frac{1}{4}\|[X,Y]\|^2. \qquad \square$$

On a matrix Lie group equipped with a metric we have three different notions of exponential maps, namely the Lie exponential map, the Riemannian exponential map and the exponential of matrices. We will now show that if we consider the Riemannian exponential map at the identity element, these three concepts agree whenever the metric is bi-invariant.

Proposition C.7 (Equivalence of Riemannian and Lie exponential). *Let G be a Lie group equipped with a bi-invariant metric. Then, the Riemannian exponential at the identity $\exp_e$ and the Lie exponential $\exp$ agree.*

Proof. Fix a vector $X_e \in \mathfrak{g}$ and consider the curve $\gamma(t) = \exp(tX_e)$. This curve is the integral curve of the invariant vector field defined by X_e, this is $\gamma'(t) = X(\gamma(t))$. For this reason, by Proposition C.6

$$\nabla_{\gamma'}\gamma' = \frac{1}{2}[X,X] = 0.$$

so $\exp(tX_e)$ is a geodesic and the result readily follows. $\square$

Proposition C.8 (Equivalence of Lie exponential and exponential of matrices). *Let G be a matrix Lie group, that is, a closed subgroup of $\mathrm{GL}(n,\mathbb{C})$. Then the matrix exponential $\exp_M$ and the Lie exponential $\exp$ agree.*

Proof. The matrix exponential $\exp_{tX}$ can be expressed as the solution of the matrix differential equation

$$\gamma'(t) = X\gamma(t) \qquad \gamma(0) = \mathrm{I},\ t \in \mathbb{R}$$

for $X \in \mathbb{C}^{n\times n} = \mathfrak{gl}(n,\mathbb{C})$. This is exactly the equation that defines the Lie exponential map as the integral curve of a right-invariant vector field, that is, the Lie exponential. $\square$

Finally, all these equivalences give a short proof of the fact that the Lie exponential map is surjective on a connected Lie group with a bi-invariant metric.

Theorem C.9 (Lie exponential surjectivity). *Let G be a connected Lie group equipped with a bi-invariant metric. The Lie exponential is surjective.*

Proof. As the Lie exponential is defined in the whole Lie algebra, so is the map $\exp_e$. Since the metric is bi-invariant, we have that at a point $(g,v) \in TG$, $\exp_g(v) = L_g(\exp_e((\mathrm{d}L_{g^{-1}})_g(v)))$ and since left-translations are diffeomorphisms, the Riemannian exponential is defined in the whole tangent bundle. Therefore, by the Hopf-Rinow theorem, this implies that there exists a geodesic between any two points. Since the geodesics starting at the identity are given by the curves $\gamma(t) = \exp(tX_e)$ for $X_e \in \mathfrak{g}$, the result follows. $\square$

Now that we have all the necessary tools, we shall return to the problem of studying the metric induced by the exponential parametrization.

As we have seen, the problem that we are interested in solving is

$$\min_{g\in G} f(g),$$

where G is a matrix Lie group equipped with an bi-invariant metric. The exponential parametrization maps this problem back to the Lie algebra

$$\min_{X\in\mathfrak{g}} f(\exp(X)).$$

Since $\mathfrak{g}$ is a vector space, putting a basis on it we have that we can use all the classical toolbox developed for Euclidean spaces to approach a solution for this problem. In particular, in the context of neural networks, we are interested in studying first-order optimization methods that approach a solution to this problem, in particular, gradient descent methods. The gradient descent update step for this problem with learning rate $\eta > 0$ is given by

$$X \leftarrow X - \eta \nabla (f \circ \exp)(X),$$

where the gradient is defined with respect to the metric, that is, it is the vector such that

$$\mathrm{d}(f \circ \exp)_X(Y) = \langle \nabla (f \circ \exp)(X), Y \rangle.$$

To study this optimization method, we first have to make sense of the gradient $\nabla (f \circ \exp)(X)$. To do so, we will make use of the differential of the exponential map.

Proposition C.10. *The differential of the exponential map on a matrix Lie group is given by the formula*

$$(\mathrm{d}\exp)_X(Y) = e^X \sum_{k=0}^{\infty} \frac{(-\operatorname{ad}_X)^k}{(k+1)!}(Y) \qquad \forall X, Y \in \mathfrak{g}.$$

Proof. (Hall, 2015) Theorem 5.4. □

An analogous formula still holds in the general case, but the proof is more delicate. The powers of the adjoint representation are to be thought as the composition of endomorphisms on $\mathfrak{g}$. For this reason, this formula can be also expressed as

$$(\mathrm{d}\exp)_X(Y) = e^X (Y - \frac{1}{2}[X, Y] + \frac{1}{6}[X, [X, Y]] - \ldots).$$

Yet another way of looking at this expression is by defining the function

$$\begin{aligned} \phi \colon \operatorname{End}(\mathfrak{g}) &\to \operatorname{End}(\mathfrak{g}) \\ X &\mapsto \frac{1 - e^{-X}}{X} \end{aligned}$$

so that

$$(\mathrm{d}\exp)_X = \mathrm{d}L_{e^X} \circ \phi(\operatorname{ad}_X).$$

In this case, the fraction that defines ϕ is just a formal expression to refer to the formal series defined in Proposition C.10.

From this we can compute the gradient of $f \circ \exp$.

Proposition C.11. *Let $f \colon G \to \mathbb{R}$ be a function defined on a matrix Lie group equipped with a bi-invariant metric. For a matrix $A \in \mathfrak{g}$ let $B = e^A$. We have*

$$\nabla (f \circ \exp)(A) = B(\mathrm{d}\exp)_{-A}(B^{-1} \nabla f(B)) = \sum_{k=0}^{\infty} \frac{(\operatorname{ad}_A)^k}{(k+1)!}(e^{-A} \nabla f(B)).$$

Proof. Let $U \in \mathfrak{g}$. By the chain rule, we have

$$(\mathrm{d}(f \circ \exp))_A(U) = (\mathrm{d}f)_B \circ (\mathrm{d}\exp)_A(U).$$

In terms of the gradient of f with respect to the metric this is equivalent to

$$\begin{aligned} (\mathrm{d}(f \circ \exp))_A(U) &= \langle \nabla f(B), (\mathrm{d}\exp)_A(U) \rangle \\ &= \langle (\mathrm{d}\exp)_A^*(\nabla f(B)), U \rangle \end{aligned}$$

which gives

$$\nabla (f \circ \exp)(A) = (\mathrm{d}\exp)_A^*(\nabla f(B)).$$

Now we just have to compute the adjoint of the differential of the exponential function. This is now simple since

$$\begin{aligned}(\mathrm{d}\exp)_A^* &= (\mathrm{d}L_{e^A} \circ \phi(\mathrm{ad}_A))^* \\ &= \phi(\mathrm{ad}_A)^* \circ \mathrm{d}L_{e^{-A}} \\ &= \phi(\mathrm{ad}_A^*) \circ \mathrm{d}L_{e^{-A}} \\ &= \phi(\mathrm{ad}_{-A}) \circ \mathrm{d}L_{e^{-A}},\end{aligned}$$

where the second equality follows from the product being left-invariant, the third one from ϕ being analytic and the last one from Lemma C.5. □

Now we can explicitly define the update rule for the exponential parametrization

$$\begin{aligned}&\widehat{r}\colon TG \to G \\ &(e^A, U) \mapsto \exp(A + \phi(\mathrm{ad}_{-A})(e^{-A}U))^{\cdot}\end{aligned}$$

We can then study the gradient flow induced by the exponential parametrization by means of $\widehat{r}$. If $\widehat{r}$ were a retraction, then the flow induced by the exponential parametrization would have similar properties as that of Riemannian gradient descent, as shown in (Boumal et al., 2016). It turns out that the exponential parametrization induces a different flow.

Theorem C.12. *Let G be a connected matrix Lie group equipped with a bi-invariant metric. The function $\widehat{r}$ is a retraction if and only if G is abelian.*

Proof. It is clear that $\widehat{r}_g(0) = g$ for every $g \in G$. Let $A \in \mathfrak{g}$ and $B = e^A$ and let $U \in T_B G$. By the chain rule we have that

$$(\mathrm{d}\widehat{r}_B)_0(U) = (\mathrm{d}\exp)_A((\mathrm{d}\exp)_A^*(U)).$$

The map $\widehat{r}$ is a retraction if and only if

$$(\mathrm{d}\exp)_A((\mathrm{d}\exp)_A^*(U)) = U$$

holds for every $U \in T_B G$. This is equivalent to having

$$\langle (\mathrm{d}\exp)_A((\mathrm{d}\exp)_A^*(U)), H\rangle = \langle U, H\rangle$$

for every $H \in T_B G$. Taking adjoints and since the metric is left-invariant, using the formula for the adjoint of the differential of the exponential map computed in Proposition C.11, or equivalently

$$\langle (\mathrm{d}\exp)_{-A}(X), (\mathrm{d}\exp)_{-A}(Y)\rangle = \langle X, Y\rangle \qquad \forall X, Y \in \mathfrak{g}.$$

In other words, $\widehat{r}$ is a retraction if and only if the Lie exponential map is a local isometry.

Now, the Lie exponential maps $\mathfrak{g}$ into G, but $\mathfrak{g}$ equipped with its metric is flat, so it has constant sectional curvature zero. On the other hand, the sectional curvature of G is given by $\kappa(X, Y) = \frac{1}{4}\|[X, Y]\|^2$. Recall that a Lie group is abelian if and only if its Lie bracket is zero.

If the Lie bracket is zero, $(\mathrm{d}\exp)_A = (\mathrm{d}L_{e^A})_e$, and it is an isometry.

Conversely, if it is an isometry, it preserves the sectional curvature, so the Lie bracket has to be identically zero, hence the group is Abelian. □

In the abelian case, we do not only have that $\widehat{r}$ is a retraction, but also that the update rule for the exponential parametrization agrees with that of Riemannian gradient descent. Recall that the Riemannian gradient descent rule for a gradient U and a step-size η is given by

$$e^A e^{-\eta e^{-A} U}.$$

On an abelian group we have that $e^X e^Y = e^{X+Y}$. Furthermore, since the adjoint representation is zero, $(\mathrm{d}\exp)_A(U) = e^A U$. Putting these two things together we have

$$\widehat{r}(e^A, -\eta U) = e^{A - \eta e^{-A} U} = e^A e^{-\eta e^{-A} U}.$$

D. Maximal Normal Neighborhood of the Identity

Definition D.1 (Normal Neighborhood). Let $V \subseteq T_p\mathcal{M}$ be a neighborhood of 0 such that the Riemannian exponential map $\exp_p$ is a diffeomorphism. Then we say that $\exp_p V$ is a *normal neighborhood* of p.

Given that on a matrix Lie group $(\mathrm{d}\exp)_{\mathrm{I}} = \mathrm{Id}$, by the inverse function theorem, there exists a normal neighborhood around the identity matrix. In this section we will prove that the maximal open normal neighborhood of $\mathrm{SO}(n)$ (resp. $\mathrm{U}(n)$) covers almost all the group. By almost all the group we mean that the closure of the normal neighborhood is equal to the group. We do so by studying at which points we have that the map $\exp$ is no longer an immersion, or in other words, we look at the points $A \in \mathfrak{g}$ at which $\det((\mathrm{d}\exp)_A) = 0$. We will prove so for the group $\mathrm{GL}(n, \mathbb{C})$, so that the arguments readily generalize to any matrix Lie group.

Recall the definition of the matrix-valued function ϕ defined on the space of endomorphisms of the Lie algebra of $\mathrm{GL}(n, \mathbb{C})$. Specifically, since $\mathfrak{gl}(n, \mathbb{C}) \cong \mathbb{C}^{n\times n}$, we have that $\mathrm{End}(\mathfrak{gl}(n, \mathbb{C})) \cong \mathbb{C}^{n^2\times n^2}$

$$\begin{aligned}\phi\colon \mathbb{C}^{n^2\times n^2} &\to \mathbb{C}^{n^2\times n^2}\\ A &\mapsto \frac{1-e^{-A}}{A} = \sum_{k=0}^{\infty} \frac{(-A)^k}{(k+1)!}\,.\end{aligned}$$

Using this function, we can factorize the differential of the exponential function on a matrix Lie group as

$$(\mathrm{d}\exp)_A = e^A \phi(\mathrm{ad}_A).$$

Let us now compute the maximal normal neighborhood of the identity. This result is classic, but the proof here is a simplification of the classical one using an approximation argument.

Theorem D.2. *Let G be a compact and connected matrix Lie group. The exponential function is analytic, with analytic inverse on a bounded open neighborhood of the origin given by*

$$U = \{A \in \mathfrak{g} \mid |\mathrm{Im}(\lambda_i(A))| < \pi\}.$$

Proof. Given that L_{e^A} is a diffemorphism, we are interested in studying when the matrix defined by the function $\phi(\mathrm{ad}_A)$ stops being full-rank and when is it injective.

First, note that if the eigenvalues of A are λ_i, then the eigenvalues of $g(A)$ with g a complex analytic function well-defined on $\{\lambda_i\}$ are $\{g(\lambda_i)\}$. This is clear for diagonalizable matrices. Since these are dense in $\mathbb{C}^{n\times n}$, given that eigenvalues are continuous functions of the matrix, it readily generalizes to arbitrary matrices.

Let $A \in \mathbb{C}^{n^2\times n^2}$ and let $\lambda_{i,j}$ for $1 \leq i, j \leq n$ be its eigenvalues. Then, ϕ is non-singular when $\phi(\lambda_{i,j}) \neq 0$ for every $\lambda_{i,j}$. Equivalently, when $\lambda_{i,j} \neq 2\pi k i$ for $k \in \mathbb{Z}\backslash\{0\}$.

Let us now compute the eigenvalues of ad_A using the same trick as above. Let $A \in \mathbb{C}^{n\times n}$ and suppose that it is diagonalizable with eigenvalues $\{\lambda_i\}$. Let u_i be the eigenvectors of A and v_i the eigenvectors of $A^\intercal$—which is also diagonalizable with the same eigenvalues—. Since

$$\mathrm{ad}_A(u_i \otimes v_j) = (\lambda_i - \lambda_j) u_i \otimes v_j$$

we have that $\{u_i \otimes v_j\}$ are the eigenvectors of ad_A with eigenvalues $\lambda_{i,j} := \lambda_i - \lambda_j$. Now, using the same continuity argument as above have that these are the eigenvalues of ad_A for every $A \in \mathbb{C}^{n\times n}$.

From all this, we draw that $(\mathrm{d}\exp)_A$ is singular whenever A has two eigenvalues that differ by a non-zero integer multiple of $2\pi i$.

Finally, on a compact matrix Lie group every matrix is diagonalizable, so the exponential acts on the eigenvalues, but the complex variable function e^z is injective on $\{z \in \mathbb{C} \mid |\mathrm{Im}(z)| < \pi\}$, so the Lie exponential is injective on this domain as well. □

Let us look at the particular cases that we are interested in. If we set $G = \mathrm{SO}(n)$, we have that its Lie algebra are the skew-symmetric matrices. Skew-symmetric matrices have purely imaginary eigenvalues. Furthermore, since they are real

matrices, their eigenvalues come in conjugate pairs. As such, we have that the exponential map is singular on every matrix in the boundary of the set U defined in Theorem D.2.

Special orthogonal matrices are those matrices which are similar to block-diagonal matrices with diagonal blocks of the form

$$B = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

for $\theta \in (-\pi, \pi]$. On $\mathrm{SO}(2n+1)$, there is an extra block with a single 1.

Similarly, skew-symmetric matrices are those matrices which are similar to block-diagonal matrices with diagonal blocks of the form

$$A = \begin{pmatrix} 0 & \theta \\ -\theta & 0 \end{pmatrix}.$$

On $\mathfrak{so}(2n+1)$ there is an extra block with a single zero.

This outlines an elementary proof of the fact that the Lie exponential is surjective on $\mathrm{SO}(n)$ and $\mathrm{U}(n)$.

In both cases this shows that the boundary of U has measure zero and that $f(\overline{U}) = G$.

Remark. The reader familiar with Lie group theory will have noticed that this proof is exactly the standard one for the surjectivity of the exponential map using the Torus theorem, where one proves that all the maximal tori in a compact Lie group are conjugated and that every element of the group lies in some maximal torus, arriving then to the same conclusion but in much more generality.

E. Hyperparameters for the Experiments

The batch size across all the experiments was 128. The learning rates for the orthogonal parameters are 10 times less those for the non-orthogonal parameters. We fixed the seed of both Numpy and Pytorch to be 5544 for all the experiments for reproducibility. This is the same seed that was used in the experiments in (Helfrich et al., 2018). In Table 3 we refer to the optimizer and learning rate for the non-orthogonal part of the neural network simply as optimizer and learning rate.

Table 3. Hyperparameters for the Experiments in Section 5.

Dataset	Size	Optimizer	Learning Rate	Orthogonal optimizer	Orthogonal Learning Rate
Copying Problem $L = 1000$	190	RMSPROP	$2 \cdot 10^{-4}$	RMSPROP	$2 \cdot 10^{-5}$
Copying Problem $L = 2000$			$2 \cdot 10^{-4}$		$2 \cdot 10^{-5}$
MNIST	170	RMSPROP	$7 \cdot 10^{-4}$	RMSPROP	$7 \cdot 10^{-5}$
	360		$5 \cdot 10^{-4}$		$5 \cdot 10^{-5}$
	512		$3 \cdot 10^{-4}$		$3 \cdot 10^{-5}$
P-MNIST	170		10^{-3}		10^{-4}
	360		$7 \cdot 10^{-4}$		$7 \cdot 10^{-5}$
	512		$5 \cdot 10^{-4}$		$5 \cdot 10^{-5}$
TIMIT	224	ADAM	10^{-3}	RMSPROP	10^{-4}
	322		$7 \cdot 10^{-4}$		$7 \cdot 10^{-5}$
	425		$7 \cdot 10^{-4}$		$7 \cdot 10^{-5}$