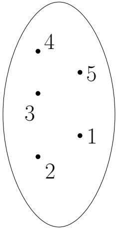
# Learning Distributions using Lie Groups

E. Mehmet Kıral

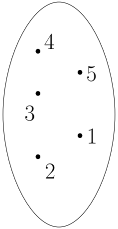
August 2025, IJCAI tutorial
AI meets Algebra[1]

[1] https://algebra4ai.github.io

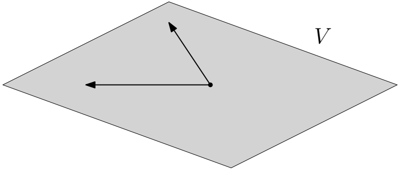
# Objects




Collection of $n$ letters.

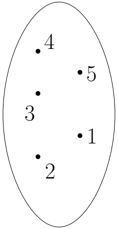



Collection of $n$ letters.

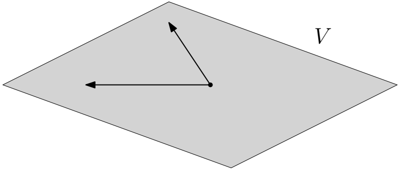



Vector Spaces

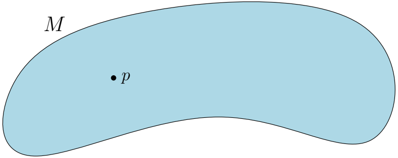
# Objects




Collection of $n$ letters.

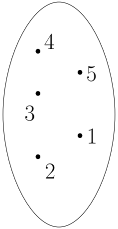



Vector Spaces

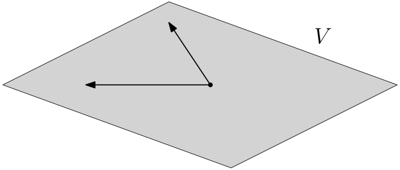



Manifolds

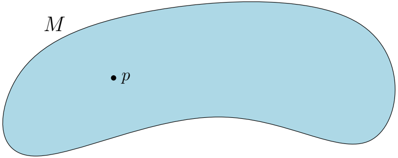
# Objects

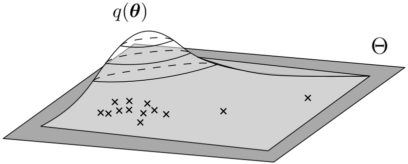



Collection of $n$ letters.




Vector Spaces



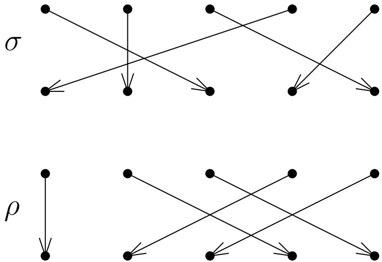




Manifolds

Distributions on a space $\Theta$.

# Transformations of objects. Groups, arise!

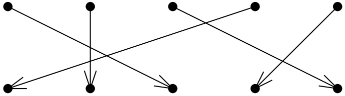
Permutations on $n$ letters. This preserves the distinctness of elements of the set.
Given two permutations you can apply one after the other.



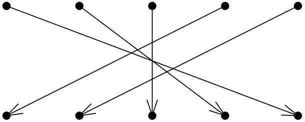
# Transformations of objects. Groups, arise!

Permutations on $n$ letters. This preserves the distinctness of elements of the set.
Given two permutations you can apply one after the other.

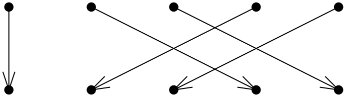
$\sigma$



$\rho \circ \sigma$

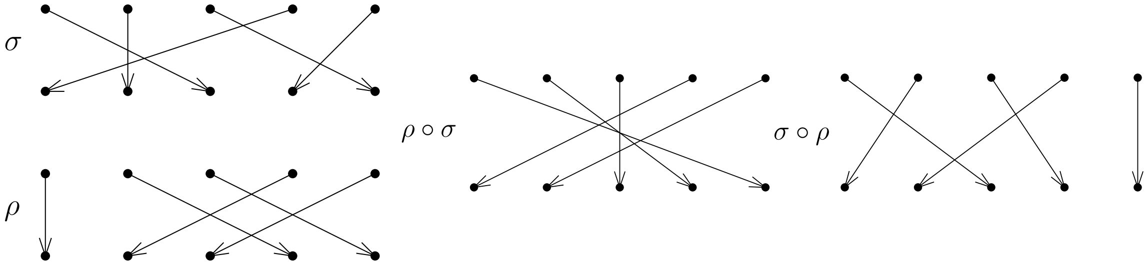


$\rho$



# Transformations of objects. Groups, arise!

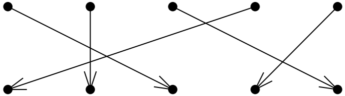
Permutations on $n$ letters. This preserves the distinctness of elements of the set.
Given two permutations you can apply one after the other.



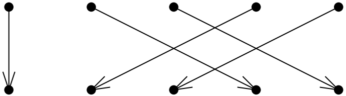
# Transformations of objects. Groups, arise!

Permutations on $n$ letters. This preserves the distinctness of elements of the set.
Given two permutations you can apply one after the other.

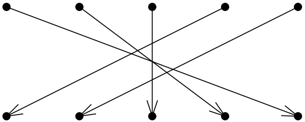
$\sigma$



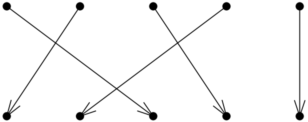
$\rho$



$\rho \circ \sigma$

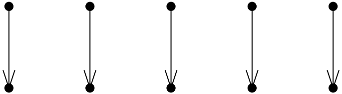


$\sigma \circ \rho$



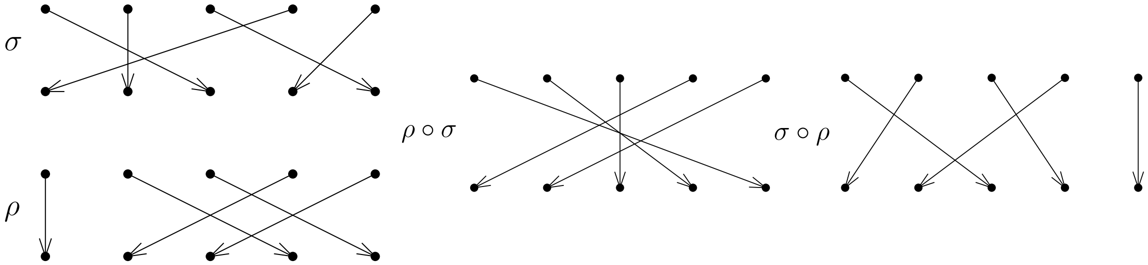
The identity permutation:

$e$

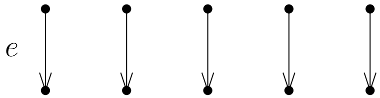


# Transformations of objects. Groups, arise!

Permutations on $n$ letters. This preserves the distinctness of elements of the set.
Given two permutations you can apply one after the other.




The identity permutation:




The resulting object is the symmetric group on $n$ letters, denoted by $S_n$ or $\mathrm{Sym}_n$.

# Matrix Groups

Invertible linear maps also preserve the linear structure of vectors. It is a group.

$$\mathrm{GL}(V) = \left\{ A : V \to V \,\middle|\, \begin{matrix} A(\mathbf{v}+\mathbf{w}) = A\mathbf{v} + A\mathbf{w}, \\ A(\alpha\mathbf{v}) = \alpha A\mathbf{v} \end{matrix}, A^{-1} \text{ exists} \right\}.$$

Also denoted by $\mathrm{GL}(n, \mathbb{R})$ if $V$ is an $n$-dimensional real vector space, i.e. $V \cong \mathbb{R}^n$.

## Matrix Groups

Invertible linear maps also preserve the linear structure of vectors. It is a group.

$$\mathrm{GL}(V) = \left\{ A : V \to V \,\middle|\, \begin{matrix} A(\mathbf{v}+\mathbf{w}) = A\mathbf{v} + A\mathbf{w}, \\ A(\alpha\mathbf{v}) = \alpha A\mathbf{v} \end{matrix}, A^{-1} \text{ exists} \right\}.$$

Also denoted by $\mathrm{GL}(n, \mathbb{R})$ if $V$ is an $n$-dimensional real vector space, i.e. $V \cong \mathbb{R}^n$.

If preserve extra structure such as inner product $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^\top \mathbf{w}$ then obtain a subgroup

$$\mathrm{SO}(n, \mathbb{R}) = \{ A \in \mathrm{GL}(n, \mathbb{R}) | \langle A\mathbf{v}, A\mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{w} \rangle, \det(A) > 0 \}.$$

# Matrix Groups

Invertible linear maps also preserve the linear structure of vectors. It is a group.

$$\mathrm{GL}(V) = \left\{ A : V \to V \,\middle|\, \begin{matrix} A(\mathbf{v}+\mathbf{w}) = A\mathbf{v} + A\mathbf{w}, \\ A(\alpha\mathbf{v}) = \alpha A\mathbf{v} \end{matrix}, A^{-1} \text{ exists} \right\}.$$

Also denoted by $\mathrm{GL}(n, \mathbb{R})$ if $V$ is an $n$-dimensional real vector space, i.e. $V \cong \mathbb{R}^n$.

If preserve extra structure such as inner product $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^\top \mathbf{w}$ then obtain a subgroup

$$\mathrm{SO}(n, \mathbb{R}) = \{ A \in \mathrm{GL}(n, \mathbb{R}) | \langle A\mathbf{v}, A\mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{w} \rangle, \det(A) > 0 \}.$$

The upper triangular matrices are also closed under multiplication and inversion

$$B = \begin{pmatrix} * & * & * & \cdots & * \\ 0 & * & * & & * \\ 0 & 0 & \ddots & & * \\ 0 & 0 & \cdots & 0 & * \end{pmatrix}$$

# Matrix Groups

Invertible linear maps also preserve the linear structure of vectors. It is a group.

$$\mathrm{GL}(V) = \left\{ A : V \to V \,\middle|\, \begin{matrix} A(\mathbf{v}+\mathbf{w}) = A\mathbf{v} + A\mathbf{w}, \\ A(\alpha\mathbf{v}) = \alpha A\mathbf{v} \end{matrix}, A^{-1} \text{ exists} \right\}.$$

Also denoted by $\mathrm{GL}(n,\mathbb{R})$ if $V$ is an $n$-dimensional real vector space, i.e. $V \cong \mathbb{R}^n$.

If preserve extra structure such as inner product $\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^\top \mathbf{w}$ then obtain a subgroup

$$\mathrm{SO}(n,\mathbb{R}) = \{A \in \mathrm{GL}(n,\mathbb{R}) | \langle A\mathbf{v}, A\mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{w} \rangle, \det(A) > 0\}.$$

The upper triangular matrices are also closed under multiplication and inversion

$$B = \begin{pmatrix} * & * & * & \cdots & * \\ 0 & * & * & & * \\ 0 & 0 & \ddots & & * \\ 0 & 0 & \cdots & 0 & * \end{pmatrix}$$

What does this set of matrices preserve?

A group is a set $G$ together with a binary operation

$$\begin{aligned} * : G \times G &\longrightarrow G \\ (g, h) &\longmapsto g * h \end{aligned}$$

satisfying

1. (associativity) $(g * h) * k = g * (h * k)$ for every $g, h, k \in G$.
2. (existence of identity) There is an $e \in G$ such that $e * g = g$ and $g * e = g$ for every $g \in G$.
3. (existence of inverses) For every $g \in G$ there exists an $h \in G$ such that $g * h = h * g = e$.

# Groups

A group is a set $G$ together with a binary operation

$$\begin{aligned} * : G \times G &\longrightarrow G \\ (g, h) &\longmapsto g * h \end{aligned}$$

satisfying

1. (associativity) $(g * h) * k = g * (h * k)$ for every $g, h, k \in G$.
2. (existence of identity) There is an $e \in G$ such that $e * g = g$ and $g * e = g$ for every $g \in G$.
3. (existence of inverses) For every $g \in G$ there exists an $h \in G$ such that $g * h = h * g = e$.

Additive symbol '$*$' $=$ '$+$' for Abelian groups: $a + b = b + a$.
Multiplicative notation mostly omits the $*$ symbol, and $ab \neq ba$ can happen (think matrices).

# Groups

A group is a set $G$ together with a binary operation

$$\begin{aligned} * : G \times G &\longrightarrow G \\ (g, h) &\longmapsto g * h \end{aligned}$$

satisfying

1. (associativity) $(g * h) * k = g * (h * k)$ for every $g, h, k \in G$.
2. (existence of identity) There is an $e \in G$ such that $e * g = g$ and $g * e = g$ for every $g \in G$.
3. (existence of inverses) For every $g \in G$ there exists an $h \in G$ such that $g * h = h * g = e$.

Additive symbol '$*$' = '$+$' for Abelian groups: $a + b = b + a$.
Multiplicative notation mostly omits the $*$ symbol, and $ab \neq ba$ can happen (think matrices).

The *closure property of the group operation* is simply that $a * b \in G$ if $a, b \in G$.

## Examples of Groups

1. $(V, +)$. Any vector space $V$ with vector addition as the binary operation.

# Examples of Groups

1. $(V, +)$. Any vector space $V$ with vector addition as the binary operation.
2. $(\mathbb{R}_{>0}, \times)$. Positive real numbers $\mathbb{R}_{>0}$ under multiplication.

# Examples of Groups

1. $(V, +)$. Any vector space $V$ with vector addition as the binary operation.
2. $(\mathbb{R}_{>0}, \times)$. Positive real numbers $\mathbb{R}_{>0}$ under multiplication.
3. $\mathrm{GL}(n, \mathbb{R})$. Invertible $n \times n$ matrices, under matrix multiplication.

# Examples of Groups

1. $(V, +)$. Any vector space $V$ with vector addition as the binary operation.
2. $(\mathbb{R}_{>0}, \times)$. Positive real numbers $\mathbb{R}_{>0}$ under multiplication.
3. $\mathrm{GL}(n, \mathbb{R})$. Invertible $n \times n$ matrices, under matrix multiplication.
4. $\mathrm{Aff}(n, \mathbb{R}) \cong \mathrm{GL}(n, \mathbb{R}) \ltimes \mathbb{R}^n$. The Affine group, consisting of pairs $(A, \mathbf{b})$ where the group multiplication is $(A_1, \mathbf{b}_1)(A_2, \mathbf{b}_2) = (A_1 A_2, A\mathbf{b}_2 + \mathbf{b}_1)$.

# Examples of Groups

1. $(V, +)$. Any vector space $V$ with vector addition as the binary operation.
2. $(\mathbb{R}_{>0}, \times)$. Positive real numbers $\mathbb{R}_{>0}$ under multiplication.
3. $\mathrm{GL}(n, \mathbb{R})$. Invertible $n \times n$ matrices, under matrix multiplication.
4. $\mathrm{Aff}(n, \mathbb{R}) \cong \mathrm{GL}(n, \mathbb{R}) \ltimes \mathbb{R}^n$. The Affine group, consisting of pairs $(A, \mathbf{b})$ where the group multiplication is $(A_1, \mathbf{b}_1)(A_2, \mathbf{b}_2) = (A_1 A_2, A\mathbf{b}_2 + \mathbf{b}_1)$.
   Another way to realize this group is as a subgroup of $(n+1) \times (n+1)$ matrices.

$$\begin{pmatrix} A_1 & \mathbf{b}_1 \\ \mathbf{0}^\top & 1 \end{pmatrix} \begin{pmatrix} A_2 & \mathbf{b}_2 \\ \mathbf{0}^\top & 1 \end{pmatrix} = \begin{pmatrix} A_1 A_2 & \mathbf{b}_1 + A_1 \mathbf{b}_2 \\ \mathbf{0}^\top & 1 \end{pmatrix} \text{ and } \begin{pmatrix} A & \mathbf{b} \\ \mathbf{0}^\top & 1 \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} & -A^{-1}\mathbf{b} \\ \mathbf{0}^\top & 1 \end{pmatrix}$$

# Examples of Groups

1. $(V, +)$. Any vector space $V$ with vector addition as the binary operation.
2. $(\mathbb{R}_{>0}, \times)$. Positive real numbers $\mathbb{R}_{>0}$ under multiplication.
3. $\mathrm{GL}(n, \mathbb{R})$. Invertible $n \times n$ matrices, under matrix multiplication.
4. $\mathrm{Aff}(n, \mathbb{R}) \cong \mathrm{GL}(n, \mathbb{R}) \ltimes \mathbb{R}^n$. The Affine group, consisting of pairs $(A, \mathbf{b})$ where the group multiplication is $(A_1, \mathbf{b}_1)(A_2, \mathbf{b}_2) = (A_1 A_2, A\mathbf{b}_2 + \mathbf{b}_1)$.
   Another way to realize this group is as a subgroup of $(n+1) \times (n+1)$ matrices.

$$\begin{pmatrix} A_1 & \mathbf{b}_1 \\ \mathbf{0}^\top & 1 \end{pmatrix} \begin{pmatrix} A_2 & \mathbf{b}_2 \\ \mathbf{0}^\top & 1 \end{pmatrix} = \begin{pmatrix} A_1 A_2 & \mathbf{b}_1 + A_1 \mathbf{b}_2 \\ \mathbf{0}^\top & 1 \end{pmatrix} \text{ and } \begin{pmatrix} A & \mathbf{b} \\ \mathbf{0}^\top & 1 \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} & -A^{-1}\mathbf{b} \\ \mathbf{0}^\top & 1 \end{pmatrix}$$

A group representation is a group homomorphism $\pi : G \to \mathrm{GL}(V)$, (i.e. $\pi(gh) = \pi(g)\pi(h)$) so that any element $g$ can be seen as a matrix $\pi(g)$.

# Examples of Groups

1. $(V, +)$. Any vector space $V$ with vector addition as the binary operation.
2. $(\mathbb{R}_{>0}, \times)$. Positive real numbers $\mathbb{R}_{>0}$ under multiplication.
3. $\mathrm{GL}(n, \mathbb{R})$. Invertible $n \times n$ matrices, under matrix multiplication.
4. $\mathrm{Aff}(n, \mathbb{R}) \cong \mathrm{GL}(n, \mathbb{R}) \ltimes \mathbb{R}^n$. The Affine group, consisting of pairs $(A, \mathbf{b})$ where the group multiplication is $(A_1, \mathbf{b}_1)(A_2, \mathbf{b}_2) = (A_1 A_2, A\mathbf{b}_2 + \mathbf{b}_1)$.
   Another way to realize this group is as a subgroup of $(n+1) \times (n+1)$ matrices.

$$\begin{pmatrix} A_1 & \mathbf{b}_1 \\ \mathbf{0}^\top & 1 \end{pmatrix} \begin{pmatrix} A_2 & \mathbf{b}_2 \\ \mathbf{0}^\top & 1 \end{pmatrix} = \begin{pmatrix} A_1 A_2 & \mathbf{b}_1 + A_1 \mathbf{b}_2 \\ \mathbf{0}^\top & 1 \end{pmatrix} \text{ and } \begin{pmatrix} A & \mathbf{b} \\ \mathbf{0}^\top & 1 \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} & -A^{-1}\mathbf{b} \\ \mathbf{0}^\top & 1 \end{pmatrix}$$

A group representation is a group homomorphism $\pi : G \to \mathrm{GL}(V)$, (i.e. $\pi(gh) = \pi(g)\pi(h)$) so that any element $g$ can be seen as a matrix $\pi(g)$.

Elements of the symmetric group $S_3$ can be represented as matrices: $\pi((12)(3)) = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

# Group Actions

A group $G$ *acts* on a set $X$ if there is a map

$$\cdot : G \times X \to X$$

satisfying

1. $(gh) \cdot x = g \cdot (h \cdot x)$ for all $g, h \in G$ and $x \in X$
2. $e \cdot x = x$ for all $x \in X$ where $e \in G$ is the identity.

# Group Actions

A group $G$ *acts* on a set $X$ if there is a map

$$\cdot : G \times X \to X$$

satisfying

1. $(gh) \cdot x = g \cdot (h \cdot x)$ for all $g, h \in G$ and $x \in X$
2. $e \cdot x = x$ for all $x \in X$ where $e \in G$ is the identity.

Every group acts on itself with the group multiplication. In other words $X = G$ and $g \cdot x = gx$.

# Group Actions

A group $G$ *acts* on a set $X$ if there is a map

$$\cdot : G \times X \to X$$

satisfying

1. $(gh) \cdot x = g \cdot (h \cdot x)$ for all $g, h \in G$ and $x \in X$
2. $e \cdot x = x$ for all $x \in X$ where $e \in G$ is the identity.

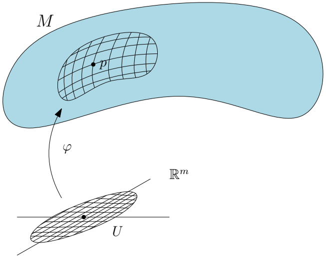
Every group acts on itself with the group multiplication. In other words $X = G$ and $g \cdot x = gx$.

Another example, $(A, \mathbf{b}) \in \mathrm{Aff}(n, \mathbb{R})$ acting on $\mathbf{x} \in \mathbb{R}^n$ by

$$(A, \mathbf{b}) \cdot \mathbf{x} = A\mathbf{x} + \mathbf{b}.$$

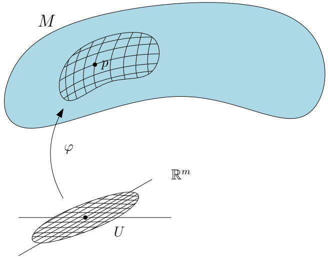
# A bit on manifolds

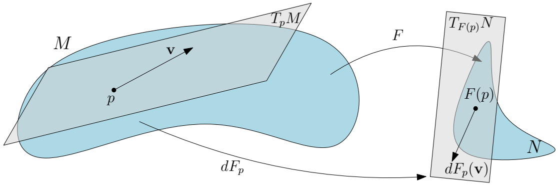
Manifolds are geometric objects which have local coordinates via charts $\varphi : U \subseteq \mathbb{R}^m \to M$.



# A bit on manifolds

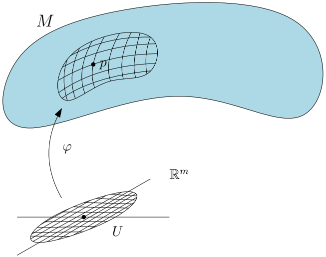
Manifolds are geometric objects which have local coordinates via charts $\varphi : U \subseteq \mathbb{R}^m \to M$.

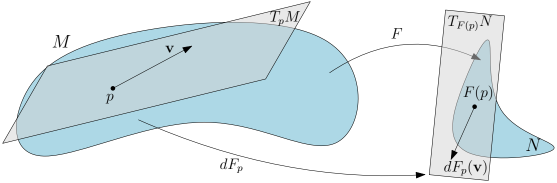



Differential of maps between two manifolds are linear maps between their tangent spaces.

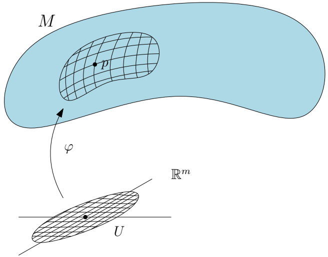
Manifolds are geometric objects which have local coordinates via charts $\varphi : U \subseteq \mathbb{R}^m \to M$.

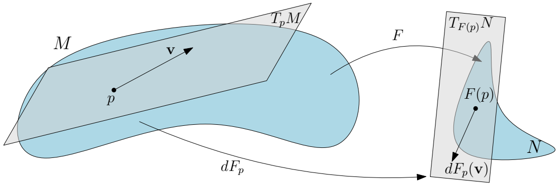



Differential of maps between two manifolds are linear maps between their tangent spaces. So if $N \subseteq \mathbb{R}$—as $T_{f(p)}N \cong \mathbb{R}$ for any point $p$—we get a linear map $df_p : T_pM \to \mathbb{R}$, i.e. a linear functional: $df_p \in T_p^*M$.

Manifolds are geometric objects which have local coordinates via charts $\varphi : U \subseteq \mathbb{R}^m \to M$.



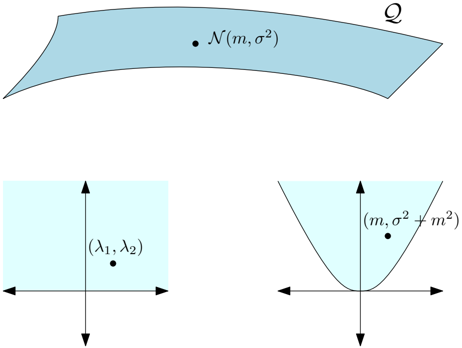

Differential of maps between two manifolds are linear maps between their tangent spaces. So if $N \subseteq \mathbb{R}$—as $T_{f(p)}N \cong \mathbb{R}$ for any point $p$—we get a linear map $df_p : T_pM \to \mathbb{R}$, i.e. a linear functional: $df_p \in T_p^*M$.

If we have two maps $F : M \to N$ and $G : N \to P$ then the chain rule is

$$d(G \circ F)_p = dG_{F(p)} dF_p$$

On the right hand side, we have a composition of two linear maps, written multiplicatively.

# Manifolds of distributions: more than one chart is possible



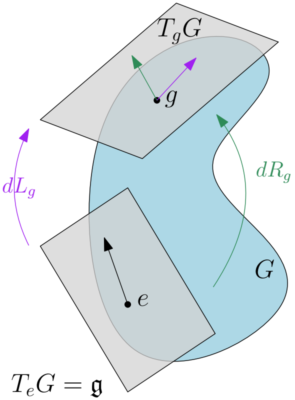

The 1-d Gaussian distributions form a two dimensional manifold with a single chart. But there can be more than one parametrization. One with natural parameters

$$(\lambda_1, \lambda_2) \longmapsto q(x) \propto e^{-\lambda_1 x - \lambda_2 x^2}$$

or on the other hand, the expectations of $x$ and $x^2$ are also enough to characterize $q(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$.

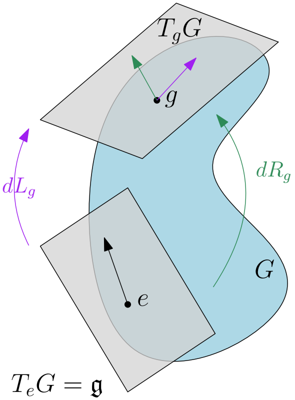
The Euclidean inner product w.r.t. $(\lambda_1, \lambda_2)$ or w.r.t. $(\mu_1, \mu_2) = (m, \sigma^2 + m^2)$ gives different angles between tangent vectors to $T_q\mathcal{Q}$.

# Lie groups: Multiplication connects tangent planes.




Lie groups are groups which are also manifolds.
Multiplication and inversion maps are smooth.

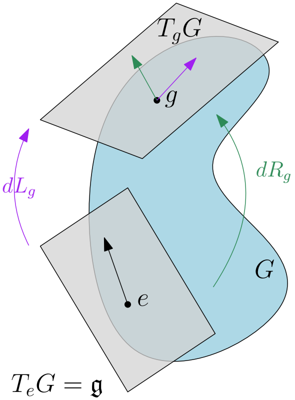
# Lie groups: Multiplication connects tangent planes.




Lie groups are groups which are also manifolds.
Multiplication and inversion maps are smooth.

For every $g \in G$ left and right multiplication maps defined by $L_g(x) = gx, R_g(x) = xg$ send $e$ to $g$.

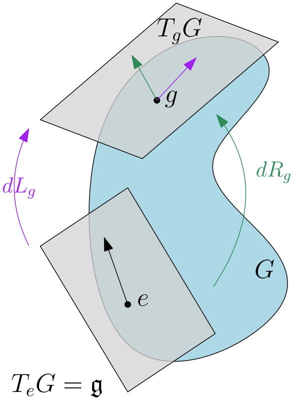
# Lie groups: Multiplication connects tangent planes.




Lie groups are groups which are also manifolds.
Multiplication and inversion maps are smooth.

For every $g \in G$ left and right multiplication maps defined by $L_g(x) = gx, R_g(x) = xg$ send $e$ to $g$.
Their differentials connect the tangent spaces $\mathfrak{g}$ and $T_gG$.

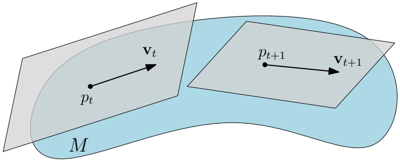
# Lie groups: Multiplication connects tangent planes.



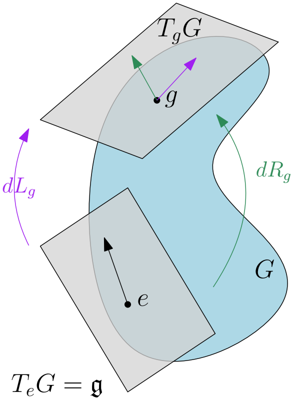

Lie groups are groups which are also manifolds.
Multiplication and inversion maps are smooth.

For every $g \in G$ left and right multiplication maps defined by $L_g(x) = gx, R_g(x) = xg$ send $e$ to $g$.
Their differentials connect the tangent spaces $\mathfrak{g}$ and $T_gG$.

If learning on a general manifold, then vectors at different points are in totally different vector spaces. How to add them, like in momentum accumulation?



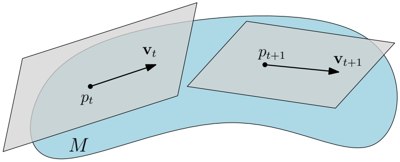
# Lie groups: Multiplication connects tangent planes.




Lie groups are groups which are also manifolds.
Multiplication and inversion maps are smooth.

For every $g \in G$ left and right multiplication maps defined by $L_g(x) = gx, R_g(x) = xg$ send $e$ to $g$.
Their differentials connect the tangent spaces $\mathfrak{g}$ and $T_gG$.

If learning on a general manifold, then vectors at different points are in totally different vector spaces. How to add them, like in momentum accumulation?

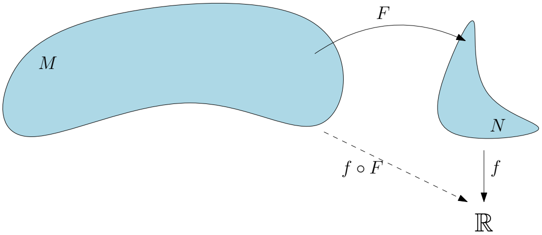



$R_{g^{-1}} \circ L_g : e \mapsto e$, its differential is the adjoint map $\mathrm{Ad}_g : \mathfrak{g} \to \mathfrak{g}$. Identity for commutative groups.

# Pushforwards and Pullbacks

Given $F : M \to N$, a map between manifolds, for every (real valued) function $f : N \to \mathbb{R}$ we can pull it back to a function $f \circ F : M \to \mathbb{R}$.

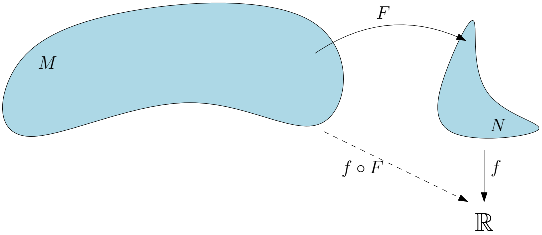
$$\begin{aligned} F^* : \mathcal{C}^\infty(N, \mathbb{R}) &\longrightarrow \mathcal{C}^\infty(M, \mathbb{R}) \\ f &\longmapsto F^*(f) = f \circ F. \end{aligned}$$



# Pushforwards and Pullbacks

Given $F : M \to N$, a map between manifolds, for every (real valued) function $f : N \to \mathbb{R}$ we can pull it back to a function $f \circ F : M \to \mathbb{R}$.

$$\begin{aligned} F^* : \mathcal{C}^\infty(N, \mathbb{R}) &\longrightarrow \mathcal{C}^\infty(M, \mathbb{R}) \\ f &\longmapsto F^*(f) = f \circ F. \end{aligned}$$



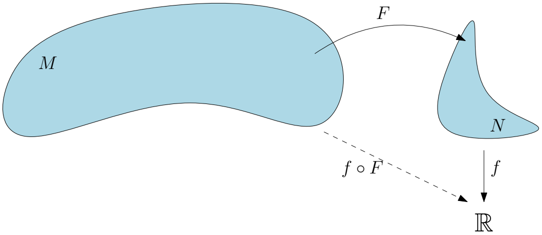

On the flip side probability distributions (or measures) $\mu$ on $M$ are pushed to $F_*\mu$ on $N$:

$$\int_N f \mathrm{d}(F_*\mu) = \int_M F^*(f) \mathrm{d}\mu.$$

# Pushforwards and Pullbacks

Given $F : M \to N$, a map between manifolds, for every (real valued) function $f : N \to \mathbb{R}$ we can pull it back to a function $f \circ F : M \to \mathbb{R}$.

$$\begin{aligned} F^* : \mathcal{C}^\infty(N, \mathbb{R}) &\longrightarrow \mathcal{C}^\infty(M, \mathbb{R}) \\ f &\longmapsto F^*(f) = f \circ F. \end{aligned}$$




On the flip side probability distributions (or measures) $\mu$ on $M$ are pushed to $F_*\mu$ on $N$:

$$\int_N f \mathrm{d}(F_*\mu) = \int_M F^*(f) \mathrm{d}\mu.$$

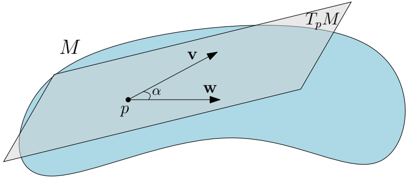
Another way to see this is on a measurable set $E \subseteq N$ the measure of the pushforward distribution is $(F^*\mu)(E) = \mu(F^{-1}(E)) = \mu(\{p \in M : F(p) \in E\})$.

# Pullbacks and pushforwards, II

The tangent vectors are pushed out to the range by the differential, a linear map

$$\mathrm{d}F_p : T_pM \to T_{F(p)}N.$$

Metrics, are bilinear forms on tangent spaces $\omega : T_pM \otimes T_pM \to \mathbb{R}$ which measure the length of the vectors and the angle between the vectors, just like the Euclidean dot product.

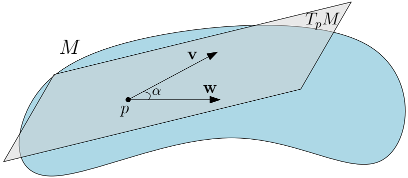


# Pullbacks and pushforwards, II

The tangent vectors are pushed out to the range by the differential, a linear map

$$\mathrm{d}F_p : T_pM \to T_{F(p)}N.$$

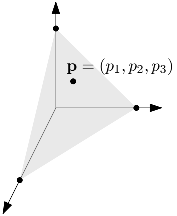
Metrics, are bilinear forms on tangent spaces $\omega : T_pM \otimes T_pM \to \mathbb{R}$ which measure the length of the vectors and the angle between the vectors, just like the Euclidean dot product.




A Riemannian manifold is a choice of bilinear form at every tangent plane $T_pM$ on $M$. The vectors go forward, and a metric $\omega$ on $N$ can be pulled back via

$$F^*(\omega)(\mathbf{v}, \mathbf{w}) = \omega(\mathrm{d}F_p\mathbf{v}, \mathrm{d}F_p\mathbf{w}).$$

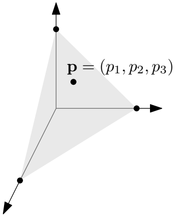
# Group acting on distributions




The group $(\mathbb{R}^r, +)$ acts on finite probability distributions on $r$ points by

$$\mathbf{x} \cdot \mathbf{p} = \mathrm{Norm}(e^{\mathbf{x}} \odot \mathbf{p}) = \left( \frac{e^{x_i} p_i}{\sum_j e^{x_j} p_j} \right)_{i=1,\dots,r}$$

# Group acting on distributions

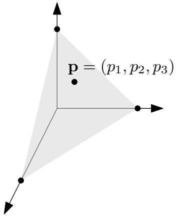



Assume $G$ acts transitively on $\boldsymbol{\theta} \in \Theta$. For example on the space of weights, e.g. $\mathbb{R}^P$.

The group $(\mathbb{R}^r, +)$ acts on finite probability distributions on $r$ points by

$$\mathbf{x} \cdot \mathbf{p} = \mathrm{Norm}(e^{\mathbf{x}} \odot \mathbf{p}) = \left( \frac{e^{x_i} p_i}{\sum_j e^{x_j} p_j} \right)_{i=1,\ldots,r}$$

# Group acting on distributions



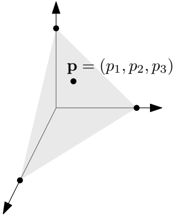

Assume $G$ acts transitively on $\boldsymbol{\theta} \in \Theta$. For example on the space of weights, e.g. $\mathbb{R}^P$.

Every $g \in G$ defines a map $\boldsymbol{\theta} \mapsto g \cdot \boldsymbol{\theta}$.

The group $(\mathbb{R}^r, +)$ acts on finite probability distributions on $r$ points by

$$\mathbf{x} \cdot \mathbf{p} = \mathrm{Norm}(e^{\mathbf{x}} \odot \mathbf{p}) = \left( \frac{e^{x_i} p_i}{\sum_j e^{x_j} p_j} \right)_{i=1,\dots,r}$$

# Group acting on distributions




The group $(\mathbb{R}^r, +)$ acts on finite probability distributions on $r$ points by

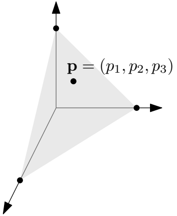
$$\mathbf{x} \cdot \mathbf{p} = \mathrm{Norm}(e^{\mathbf{x}} \odot \mathbf{p}) = \left( \frac{e^{x_i} p_i}{\sum_j e^{x_j} p_j} \right)_{i=1,\ldots,r}$$

Assume $G$ acts transitively on $\boldsymbol{\theta} \in \Theta$. For example on the space of weights, e.g. $\mathbb{R}^P$.

Every $g \in G$ defines a map $\boldsymbol{\theta} \mapsto g \cdot \boldsymbol{\theta}$.

$\nu$ a base measure, s.t. $\nu(g \cdot E) = \chi(g)\nu(E)$, e.g. $\chi(g) = \det(A)$ for $g = (A, \mathbf{b}) \in \mathrm{Aff}(n, \mathbb{R})$.

# Group acting on distributions




The group $(\mathbb{R}^r, +)$ acts on finite probability distributions on $r$ points by

$$\mathbf{x} \cdot \mathbf{p} = \mathrm{Norm}(e^{\mathbf{x}} \odot \mathbf{p}) = \left( \frac{e^{x_i} p_i}{\sum_j e^{x_j} p_j} \right)_{i=1,\ldots,r}$$

Assume $G$ acts transitively on $\boldsymbol{\theta} \in \Theta$. For example on the space of weights, e.g. $\mathbb{R}^P$.

Every $g \in G$ defines a map $\boldsymbol{\theta} \mapsto g \cdot \boldsymbol{\theta}$.

$\nu$ a base measure, s.t. $\nu(g \cdot E) = \chi(g)\nu(E)$, e.g. $\chi(g) = \det(A)$ for $g = (A, \mathbf{b}) \in \mathrm{Aff}(n, \mathbb{R})$.

The pushforward of distributions $q\mathrm{d}\nu$ are given by $g_*(q\mathrm{d}\nu) = q^g \mathrm{d}\nu$ where

$$q^g(\boldsymbol{\theta}) = \frac{1}{\chi(g)} q(g^{-1} \cdot \boldsymbol{\theta}).$$

# Parametrizing $Q$ by groups.

Assume a Lie group $G$ acts on the parameter manifold $\Theta$,

# Parametrizing $\mathcal{Q}$ by groups.

Assume a Lie group $G$ acts on the parameter manifold $\Theta$, it also acts on distributions on $\Theta$.

# Parametrizing $\mathcal{Q}$ by groups.

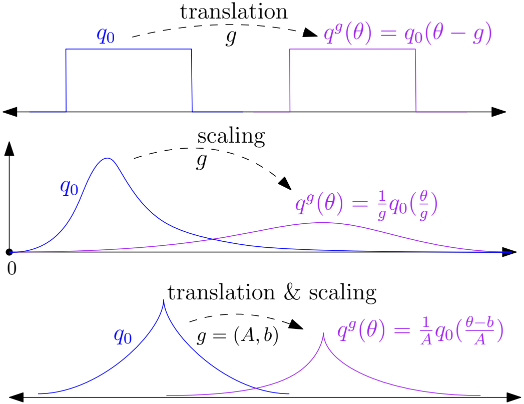
Assume a Lie group $G$ acts on the parameter manifold $\Theta$, it also acts on distributions on $\Theta$. $\mathcal{Q}$ is formed as the orbit of such an action for any base distribution $q_0$:

$$\mathcal{Q} = \{q^g \mathrm{d}\nu : g \in G\}. \qquad \text{where recall} \qquad q^g(\boldsymbol{\theta}) = \frac{1}{\chi(g)} q_0(g^{-1} \cdot \boldsymbol{\theta}).$$

# Parametrizing $\mathcal{Q}$ by groups.

Assume a Lie group $G$ acts on the parameter manifold $\Theta$, it also acts on distributions on $\Theta$. $\mathcal{Q}$ is formed as the orbit of such an action for any base distribution $q_0$:

$$\mathcal{Q} = \{q^g \mathrm{d}\nu : g \in G\}. \qquad \text{where recall} \qquad q^g(\boldsymbol{\theta}) = \frac{1}{\chi(g)} q_0(g^{-1} \cdot \boldsymbol{\theta}).$$




- $G = (\mathbb{R}, +)$, $\boldsymbol{\Theta} = \mathbb{R}$,
- $G = (\mathbb{R}_{>0}, \times)$, $\boldsymbol{\Theta} = \mathbb{R}_{>0}$,
- $G = \mathrm{Aff}(\mathbb{R}) = \mathbb{R}_{>0} \ltimes \mathbb{R}$, $\boldsymbol{\Theta} = \mathbb{R}$

# The classical and Bayesian learning setups (what we learn!)

Data

parameter space

$$\ell : \Theta \longrightarrow \mathbb{R}_{\geq 0}$$

Model

loss function

# The classical and Bayesian learning setups (what we learn!)

Data

parameter space

$$\ell : \Theta \longrightarrow \mathbb{R}_{\geq 0}$$

Model

loss function

Classically: find $\theta^* \in \Theta$ minimizing $\ell$.

Bayesian : find a distribution $q \in \mathcal{P}_\nu(\Theta)$

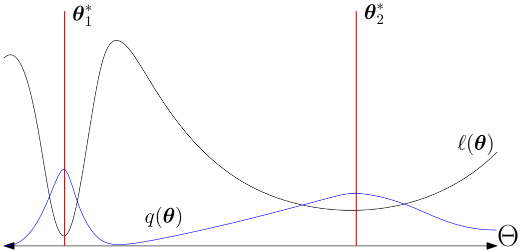
....

# Classical vs. Bayesian learning

The loss function is highly nonconvex. Usually

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}) + R(\boldsymbol{\theta})$$

where $\ell_i(\boldsymbol{\theta})$ is the loss contribution from the $i^{\text{th}}$ data point and $R(\boldsymbol{\theta})$ regularizer.




$\boldsymbol{\theta}_1^*$ and $\boldsymbol{\theta}_2^*$ are both equally valid explanations of the same data.
A distribution over the data considers both explanations "*at the same time*".

# Betting it all on one outcome

Say two dice are thrown and I tell you that the sum is greater than $7$.

# Betting it all on one outcome

Say two dice are thrown and I tell you that the sum is greater than 7.⚄, ⚂ satisfies this.

We could say the result was definitely ⚄, ⚂.

# Betting it all on one outcome

Say two dice are thrown and I tell you that the sum is greater than 7. ⚄, ⚂ satisfies this.

We could say the result was definitely ⚄, ⚂.

But there are a total of 15 possibilities

| | | | | |
|---|---|---|---|---|
| ⚅, ⚁ | ⚄, ⚂ | ⚃, ⚃ | ⚂, ⚄ | ⚁, ⚅ |
| ⚅, ⚂ | ⚄, ⚃ | ⚃, ⚄ | ⚂, ⚅ | |
| ⚅, ⚃ | ⚄, ⚄ | ⚃, ⚅ | | |
| ⚅, ⚄ | ⚄, ⚅ | | | |
| ⚅, ⚅ | | | | |

It is much more sensible to say it is one of these 15 outcomes, with equal probability.

(principle of indifference, principle of maximum entropy)

# The Bayesian Learning Problem

$\ell(\boldsymbol{\theta})$, a loss function on model parameters $\boldsymbol{\theta} \in \Theta$. Pick a base measure $\nu$ on $\Theta$.

# The Bayesian Learning Problem

$\ell(\boldsymbol{\theta})$, a loss function on model parameters $\boldsymbol{\theta} \in \Theta$. Pick a base measure $\nu$ on $\Theta$. We solve

$$q_* \in \arg\min_{q \in \mathcal{Q}} \underbrace{\mathbb{E}_q[\ell] - \tau \mathcal{H}_\nu(q)}_{=: \mathcal{E}(q)}$$

for some family of distributions $\mathcal{Q} \subseteq \mathcal{P}_\nu(\Theta) = \{q(\boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{\theta})\}$ on the parameters.

# The Bayesian Learning Problem

$\ell(\boldsymbol{\theta})$, a loss function on model parameters $\boldsymbol{\theta} \in \Theta$. Pick a base measure $\nu$ on $\Theta$. We solve

$$q_* \in \arg\min_{q \in \mathcal{Q}} \underbrace{\mathbb{E}_q[\ell] - \tau \mathcal{H}_\nu(q)}_{=:\mathcal{E}(q)}$$

for some family of distributions $\mathcal{Q} \subseteq \mathcal{P}_\nu(\Theta) = \{q(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})\}$ on the parameters.

- The expectation $\mathbb{E}_q[\ell] = \int_\Theta \ell(\boldsymbol{\theta})q(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})$ prefers regions with low loss.

# The Bayesian Learning Problem

$\ell(\boldsymbol{\theta})$, a loss function on model parameters $\boldsymbol{\theta} \in \Theta$. Pick a base measure $\nu$ on $\Theta$. We solve

$$q_* \in \arg\min_{q \in \mathcal{Q}} \underbrace{\mathbb{E}_q[\ell] - \tau \mathcal{H}_\nu(q)}_{=:\mathcal{E}(q)}$$

for some family of distributions $\mathcal{Q} \subseteq \mathcal{P}_\nu(\Theta) = \{q(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})\}$ on the parameters.

- The expectation $\mathbb{E}_q[\ell] = \int_\Theta \ell(\boldsymbol{\theta})q(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})$ prefers regions with low loss.
- The entropy $\mathcal{H}_\nu(q) = -\int_\Theta q(\boldsymbol{\theta}) \log q(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})$ prefers a higher spread of $q$.

# The Bayesian Learning Problem

$\ell(\boldsymbol{\theta})$, a loss function on model parameters $\boldsymbol{\theta} \in \Theta$. Pick a base measure $\nu$ on $\Theta$. We solve

$$q_* \in \arg\min_{q \in \mathcal{Q}} \underbrace{\mathbb{E}_q[\ell] - \tau \mathcal{H}_\nu(q)}_{=:\mathcal{E}(q)}$$

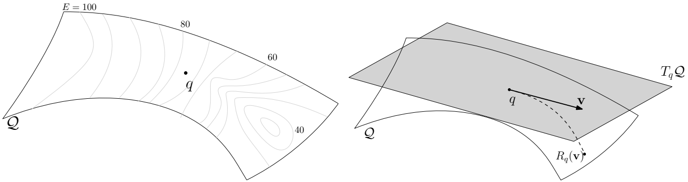
for some family of distributions $\mathcal{Q} \subseteq \mathcal{P}_\nu(\Theta) = \{q(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})\}$ on the parameters.

- The <u>expectation</u> $\mathbb{E}_q[\ell] = \int_\Theta \ell(\boldsymbol{\theta}) q(\boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{\theta})$ prefers regions with low loss.
- The <u>entropy</u> $\mathcal{H}_\nu(q) = -\int_\Theta q(\boldsymbol{\theta}) \log q(\boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{\theta})$ prefers a higher spread of $q$.
- The <u>temperature</u> $\tau > 0$ is a balancing term.

# Gradient Descent on Manifolds

Gradient descent for $\mathcal{E}(q)$ on the manifold $\mathcal{Q}$ has three components.

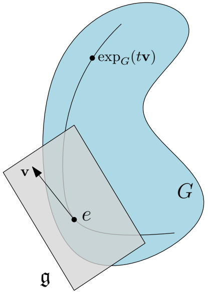
1. Calculate the differential $\mathrm{d}\mathcal{E}|_q : T_q\mathcal{Q} \mapsto \mathbb{R}$.
2. Find a way to turn the covector $\mathrm{d}\mathcal{E}|_q \in T_q^*\mathcal{Q}$ into a vector $\mathbf{v} \in T_q\mathcal{Q}$.
3. A choice of retraction brings us back onto the manifold $R_q(\mathbf{v}) \in \mathcal{Q}$.

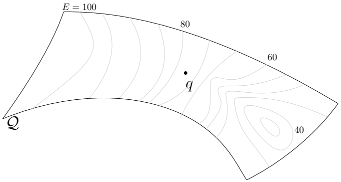


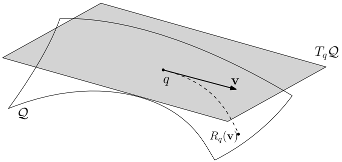
# Gradient Descent on Manifolds

Gradient descent for $\mathcal{E}(q)$ on the manifold $\mathcal{Q}$ has three components.

1. Calculate the differential $\mathrm{d}\mathcal{E}|_q : T_q\mathcal{Q} \mapsto \mathbb{R}$.
2. Find a way to turn the covector $\mathrm{d}\mathcal{E}|_q \in T_q^*\mathcal{Q}$ into a vector $\mathbf{v} \in T_q\mathcal{Q}$.
3. A choice of retraction brings us back onto the manifold $R_q(\mathbf{v}) \in \mathcal{Q}$.




On every Lie group there is an exponential map

$$\exp_G : \mathfrak{g} \to G$$

provides a canonical retraction.

# Optimization on the group: calculating the differential

We now solve

$$\arg\min_{g\in G} \mathcal{E}(q^g) = \arg\min_{g\in G} \int_\Theta q^g(\boldsymbol{\theta})\ell(\boldsymbol{\theta}) + \tau q^g(\boldsymbol{\theta}) \log q^g(\boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{\theta})$$

Given $X \in \mathfrak{g} = T_e G$ the differential in the direction of $X$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(q^{ge^{tX}})\Big|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t} \underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta})\ell(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{data contribution}} + \tau \underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta}) \log q^{ge^{tX}}(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{entropy contribution}}\Big|_{t=0}$$

# Optimization on the group: calculating the differential

We now solve

$$\arg\min_{g\in G} \mathcal{E}(q^g) = \arg\min_{g\in G} \int_\Theta q^g(\boldsymbol{\theta})\ell(\boldsymbol{\theta}) + \tau q^g(\boldsymbol{\theta}) \log q^g(\boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{\theta})$$

Given $X \in \mathfrak{g} = T_e G$ the differential in the direction of $X$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(q^{ge^{tX}})\Big|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t} \underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta})\ell(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{data contribution}} + \tau \underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta}) \log q^{ge^{tX}}(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{entropy contribution}}\Bigg|_{t=0}$$

The data contribution can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Theta q^{ge^{tX}}(\boldsymbol{\theta})\ell(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})\Big|_{t=0}$$

# Optimization on the group: calculating the differential

We now solve

$$\arg\min_{g\in G} \mathcal{E}(q^g) = \arg\min_{g\in G} \int_\Theta q^g(\boldsymbol{\theta})\ell(\boldsymbol{\theta}) + \tau q^g(\boldsymbol{\theta}) \log q^g(\boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{\theta})$$

Given $X \in \mathfrak{g} = T_e G$ the differential in the direction of $X$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(q^{ge^{tX}})\Big|_{t=0} = \underbrace{\frac{\mathrm{d}}{\mathrm{d}t}\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta})\ell(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{data contribution}} + \tau \underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta}) \log q^{ge^{tX}}(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{entropy contribution}}\Big|_{t=0}$$

The data contribution can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Theta \frac{1}{\chi(ge^{tX})} q_0((ge^{tX})^{-1} \cdot \boldsymbol{\theta})\ell(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})\Big|_{t=0}$$

# Optimization on the group: calculating the differential

We now solve

$$\arg\min_{g\in G} \mathcal{E}(q^g) = \arg\min_{g\in G} \int_\Theta q^g(\boldsymbol{\theta})\ell(\boldsymbol{\theta}) + \tau q^g(\boldsymbol{\theta}) \log q^g(\boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{\theta})$$

Given $X \in \mathfrak{g} = T_e G$ the differential in the direction of $X$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(q^{ge^{tX}})\Big|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t} \underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta})\ell(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{data contribution}} + \tau \underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta}) \log q^{ge^{tX}}(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{entropy contribution}}\Bigg|_{t=0}$$

The data contribution can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Theta \frac{1}{\chi(ge^{tX})} q_0(\boldsymbol{\theta})\ell(ge^{tX}\cdot\boldsymbol{\theta})\mathrm{d}\nu(ge^{tX}\cdot\boldsymbol{\theta})\Big|_{t=0}$$

# Optimization on the group: calculating the differential

We now solve

$$\arg\min_{g\in G} \mathcal{E}(q^g) = \arg\min_{g\in G} \int_\Theta q^g(\boldsymbol{\theta})\ell(\boldsymbol{\theta}) + \tau q^g(\boldsymbol{\theta}) \log q^g(\boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{\theta})$$

Given $X \in \mathfrak{g} = T_e G$ the differential in the direction of $X$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(q^{ge^{tX}})\Big|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t} \underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta})\ell(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{data contribution}} + \tau \underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta}) \log q^{ge^{tX}}(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{entropy contribution}}\Big|_{t=0}$$

The data contribution can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Theta q_0(\boldsymbol{\theta})\ell(ge^{tX}\cdot\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})\Big|_{t=0}$$

# Optimization on the group: calculating the differential

We now solve

$$\arg\min_{g\in G} \mathcal{E}(q^g) = \arg\min_{g\in G} \int_\Theta q^g(\boldsymbol{\theta})\ell(\boldsymbol{\theta}) + \tau q^g(\boldsymbol{\theta}) \log q^g(\boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{\theta})$$

Given $X \in \mathfrak{g} = T_e G$ the differential in the direction of $X$ is

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{E}(q^{ge^{tX}})\bigg|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t} \underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta})\ell(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{data contribution}} + \tau \underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta}) \log q^{ge^{tX}}(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{entropy contribution}}\Bigg|_{t=0}$$

The data contribution can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Theta q_0(\boldsymbol{\theta})\ell(ge^{tX}g^{-1}g \cdot \boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})\bigg|_{t=0}$$

# Optimization on the group: calculating the differential

We now solve

$$\arg\min_{g\in G} \mathcal{E}(q^g) = \arg\min_{g\in G} \int_\Theta q^g(\boldsymbol{\theta})\ell(\boldsymbol{\theta}) + \tau q^g(\boldsymbol{\theta}) \log q^g(\boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{\theta})$$

Given $X \in \mathfrak{g} = T_e G$ the differential in the direction of $X$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(q^{ge^{tX}})\bigg|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t} \underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta})\ell(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{data contribution}} + \tau \underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta}) \log q^{ge^{tX}}(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{entropy contribution}}\bigg|_{t=0}$$

The data contribution can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Theta \frac{1}{\chi(g)} q_0(g^{-1}\cdot\boldsymbol{\theta})\ell(ge^{tX}g^{-1}\cdot\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})\bigg|_{t=0}$$

# Optimization on the group: calculating the differential

We now solve

$$\arg\min_{g\in G} \mathcal{E}(q^g) = \arg\min_{g\in G} \int_\Theta q^g(\boldsymbol{\theta})\ell(\boldsymbol{\theta}) + \tau q^g(\boldsymbol{\theta}) \log q^g(\boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{\theta})$$

Given $X \in \mathfrak{g} = T_e G$ the differential in the direction of $X$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(q^{ge^{tX}})\bigg|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t} \underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta})\ell(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{data contribution}} + \tau \underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta}) \log q^{ge^{tX}}(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{entropy contribution}}\Bigg|_{t=0}$$

The data contribution can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Theta q^g(\boldsymbol{\theta})\ell(ge^{tX}g^{-1} \cdot \boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})\bigg|_{t=0}$$

# Optimization on the group: calculating the differential

We now solve

$$\arg\min_{g\in G}\mathcal{E}(q^g) = \arg\min_{g\in G}\int_\Theta q^g(\boldsymbol{\theta})\ell(\boldsymbol{\theta}) + \tau q^g(\boldsymbol{\theta})\log q^g(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})$$

Given $X \in \mathfrak{g} = T_e G$ the differential in the direction of $X$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(q^{ge^{tX}})\bigg|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t}\underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta})\ell(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{data contribution}} + \tau\underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta})\log q^{ge^{tX}}(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{entropy contribution}}\Bigg|_{t=0}$$

The data contribution can be rewritten as

$$\int_\Theta q^g(\boldsymbol{\theta})\nabla\ell(ge^{\mathbf{0}}g^{-1}\cdot\boldsymbol{\theta})^\top\frac{\mathrm{d}}{\mathrm{d}t}\left((ge^{tX}g^{-1})\cdot\boldsymbol{\theta}\right)\Big|_{t=0}\mathrm{d}\nu(\boldsymbol{\theta})$$

# Optimization on the group: calculating the differential

We now solve

$$\arg\min_{g \in G} \mathcal{E}(q^g) = \arg\min_{g \in G} \int_\Theta q^g(\boldsymbol{\theta})\ell(\boldsymbol{\theta}) + \tau q^g(\boldsymbol{\theta}) \log q^g(\boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{\theta})$$

Given $X \in \mathfrak{g} = T_e G$ the differential in the direction of $X$ is

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{E}(q^{ge^{tX}})\Big|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t} \underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta})\ell(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{data contribution}} + \tau \underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta}) \log q^{ge^{tX}}(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{entropy contribution}}\Big|_{t=0}$$

The data contribution can be rewritten as

$$\int_\Theta q^g(\boldsymbol{\theta}) \nabla \ell(\boldsymbol{\theta})^\top (\mathrm{Ad}_g(X) \cdot \boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{\theta})$$

# Optimization on the group: calculating the differential

We now solve

$$\arg\min_{g\in G} \mathcal{E}(q^g) = \arg\min_{g\in G} \int_\Theta q^g(\boldsymbol{\theta})\ell(\boldsymbol{\theta}) + \tau q^g(\boldsymbol{\theta}) \log q^g(\boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{\theta})$$

Given $X \in \mathfrak{g} = T_eG$ the differential in the direction of $X$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(q^{ge^{tX}})\Big|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t} \underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta})\ell(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{data contribution}} + \tau \underbrace{\int_\Theta q^{ge^{tX}}(\boldsymbol{\theta}) \log q^{ge^{tX}}(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})}_{\text{entropy contribution}}\Bigg|_{t=0}$$

The data contribution can be rewritten as

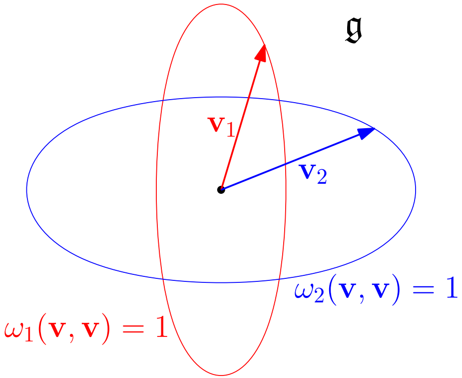
$$\int_\Theta q^g(\boldsymbol{\theta})\nabla\ell(\boldsymbol{\theta})^\top(\mathrm{Ad}_g(X)\cdot\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta}) \approx \frac{1}{K}\sum_{\substack{i=1\\ \theta_i\sim q^g}}^{K} \nabla\ell(\boldsymbol{\theta}_i)^\top(\mathrm{Ad}_g(X)\cdot\boldsymbol{\theta}_i)$$

# The metric determines the *fastest direction of descent*.

Which descent direction is *fastest* depends on how we measure distances in the tangent space.
$\mathbf{v}_1$ is fastest if we measure $\|v\|$ using $\omega_1$
$\mathbf{v}_2$ is fastest if $\|v\|$ is measured using $\omega_2$.
The pullback the Fisher information metric on $T_{q^g}\mathcal{Q}$

$$\omega^{\text{Fisher}}(h_1, h_2) = \mathbb{E}_q\left[(\partial_{h_1} \log q^g)(\partial_{h_2} \log q^g)\right]$$

to $\mathfrak{g}$ is independent of $g \in G$.



# The metric determines the *fastest direction of descent*.

Which descent direction is *fastest* depends on how we measure distances in the tangent space.

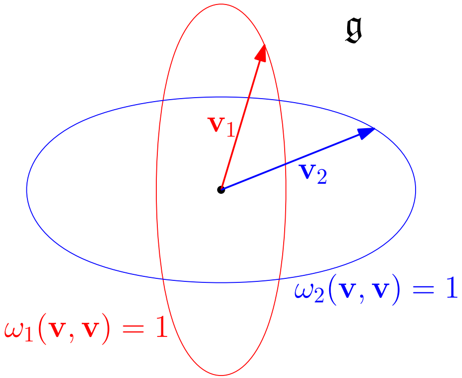
$\mathbf{v}_1$ is fastest if we measure $\|v\|$ using $\omega_1$

$\mathbf{v}_2$ is fastest if $\|v\|$ is measured using $\omega_2$.

The pullback the Fisher information metric on $T_{q^g}\mathcal{Q}$

$$\omega^{\mathsf{Fisher}}(h_1, h_2) = \mathbb{E}_q\left[(\partial_{h_1}\log q^g)(\partial_{h_2}\log q^g)\right]$$

to $\mathfrak{g}$ is independent of $g \in G$.

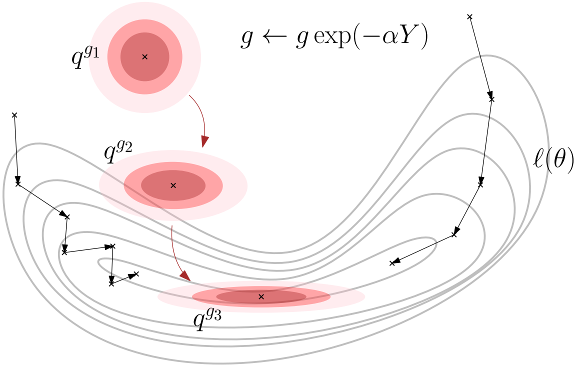



Indeed $h_X^g(\boldsymbol{\theta}) := \frac{\mathrm{d}}{\mathrm{d}t} q^{ge^{tX}}(\boldsymbol{\theta})\Big|_{t=0} = \frac{1}{\chi(g)} h_X^e(g^{-1}\cdot\boldsymbol{\theta})$, so a change of variables $\boldsymbol{\theta} \mapsto g\cdot\boldsymbol{\theta}$ shows

$$\omega^g(X, Y) = \int_\Theta \frac{h_X^g(\boldsymbol{\theta})}{q^g(\boldsymbol{\theta})}\frac{h_X^g(\boldsymbol{\theta})}{q^g(\boldsymbol{\theta})} q^g(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta}) = \omega^e(X, Y).$$

# Classical Learning vs. Learning via Group

The *point based* gradient descent updates parameters: $\theta \leftarrow \theta - \alpha \nabla \ell(\theta)$
Bayesian Learning Rule(s) update the distribution over the parameters $\theta$.



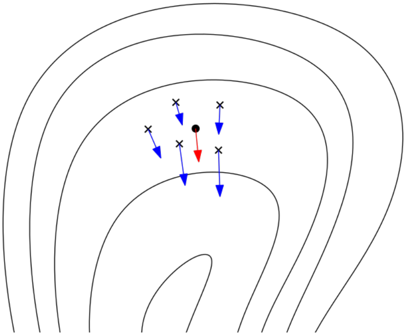

$Y \in T_e G$ is the direction of fastest ascent of $\mathcal{E}(q^g)$ w.r.t. the Fisher metric.

# Specific Update Formulas: The Additive Group

$g \in \mathbb{R}^P$ additive $\implies$ $g \longleftarrow g - \alpha \mathbb{E}_{q_g}\left[\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})\right]$

# Specific Update Formulas: The Additive Group

$g \in \mathbb{R}^P$ additive $\implies$ $g \longleftarrow g - \alpha \mathbb{E}_{q_g} \left[ \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) \right]$



Instead of going in the direction of the derivative at $g$, the direction is chosen by consensus with at points sampled from $q_g$.

# Multiplicative and Affine Update Formulas

If $g \in (\mathbb{R}^P_{>0}, \times)$ acting on a parameters $\boldsymbol{\theta} \in \mathbb{R}^P_+$:

$$X = \mathbb{E}_{q^g}[\boldsymbol{\theta} \odot \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})] - \tau$$
$$g \longleftarrow g \odot e^{-\alpha X}$$

# Multiplicative and Affine Update Formulas

If $g \in (\mathbb{R}^P_{>0}, \times)$ acting on a parameters $\boldsymbol{\theta} \in \mathbb{R}^P_+$:

$$X = \mathbb{E}_{q^g}[\boldsymbol{\theta} \odot \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})] - \tau$$
$$g \longleftarrow g \odot e^{-\alpha X}$$

For the Affine group the Lie algebra elements (tangent vectors) are of the form $\begin{pmatrix} X & \mathbf{y} \\ \mathbf{0}^\top & 0 \end{pmatrix}$ and the gradient update at $q^g$ is

$$X = \mathbb{E}_{q_0}\left[A^\top \nabla_{\boldsymbol{\theta}} \ell(A\boldsymbol{\theta} + \mathbf{b})\boldsymbol{\theta}^\top\right] - \tau I \qquad \mathbf{y} = \mathbb{E}_{q_0}\left[A^\top \nabla \ell(A\boldsymbol{\theta} + \mathbf{b})\right].$$

with a slightly more involved exponential map

$$A \longleftarrow A e^{-\alpha X} \qquad \mathbf{b} \longleftarrow A \frac{e^{-\alpha X} - I}{X} \mathbf{y} + \mathbf{b}$$

# Filters of the multiplicative group

Label nodes in a neural network "excitatory" or "inhibitory" like biology.

Magnitudes of the weights (in $\mathbb{R}_{>0}$) are the parameters (signs are fixed).
At each layer the map is $\mathbf{x} \mapsto \sigma(W_+\mathbf{x} - W_-\mathbf{x})$.

# Filters of the multiplicative group

Label nodes in a neural network "excitatory" or "inhibitory" like biology.

Magnitudes of the weights (in $\mathbb{R}_{>0}$) are the parameters (signs are fixed).
At each layer the map is $\mathbf{x} \mapsto \sigma(W_+\mathbf{x} - W_-\mathbf{x})$.

Given $g \in \mathbb{R}^P_{>0}$, and $q_0$ Rayleigh, say, and $\theta_j \sim q_0^P$ for $j = 1, \dots, K$

$$M \leftarrow \beta M + (1-\beta)\frac{1}{K}\sum_{j=1}^{K}(g \odot \boldsymbol{\theta}_j)\nabla\ell(g \odot \boldsymbol{\theta}_j) - \tau$$

$$g \leftarrow g \odot \exp\left(-\alpha M\right)$$

# Filters of the multiplicative group

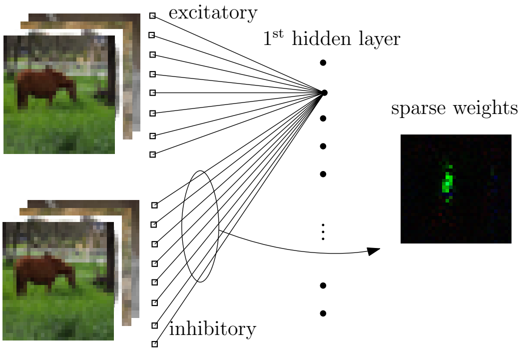
Label nodes in a neural network "excitatory" or "inhibitory" like biology.

Magnitudes of the weights (in $\mathbb{R}_{>0}$) are the parameters (signs are fixed).
At each layer the map is $\mathbf{x} \mapsto \sigma(W_+\mathbf{x} - W_-\mathbf{x})$.

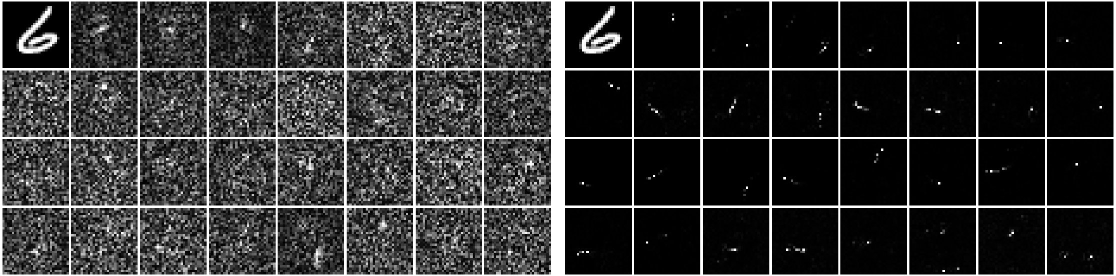
Given $g \in \mathbb{R}^P_{>0}$, and $q_0$ Rayleigh, say, and $\theta_j \sim q_0^P$ for $j = 1, \ldots, K$

$$M \leftarrow \beta M + (1-\beta)\frac{1}{K}\sum_{j=1}^{K}(g \odot \boldsymbol{\theta}_j)\nabla\ell(g \odot \boldsymbol{\theta}_j) - \tau$$

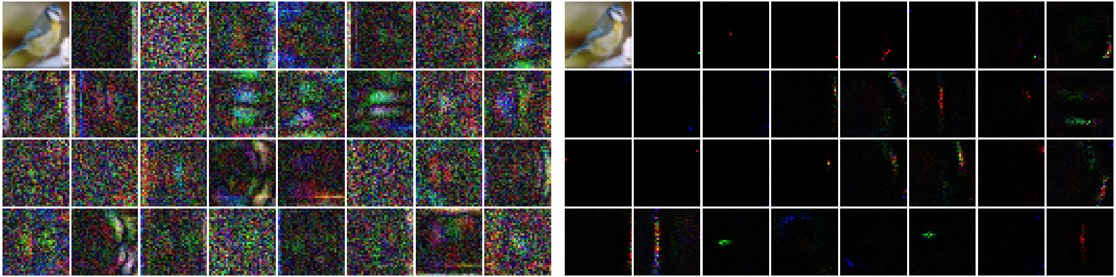
$$g \leftarrow g \odot \exp\left(-\alpha M\right)$$



# Multiplicative vs Additive filters



# The additive vs multiplicative filters for RGB images



Teşekkürler
ありがとうございます
Vielen Danke
Merci
Thank you.[2]

[2] More can be found at my blog-posts via https://ekiral.github.io/blog/blog-index.html

# Stiefel Manifold Update

Assume parameters are given as a matrix and want to preserve orthogonality of columns.

$$\Theta = \mathrm{St}(n, m) = \{\theta \in \mathrm{Mat}(n, m) : \theta^\top \theta = I_{m\times m}\}$$

The group $S = \mathrm{SO}(n)$ preserves this manifold. And given a loss function $\ell : \Theta \to \mathbb{R}_{\geq 0}$

$Y \in \mathfrak{so}(n)$ the update direction $\quad Y = \mathrm{Skew Y_0} = \frac{Y_0 - Y_0^\top}{2} \quad Y_0 = \mathbb{E}_{q_\Lambda}[\nabla \ell \theta^\top]$

# Stiefel Manifold Update

Assume parameters are given as a matrix and want to preserve orthogonality of columns.

$$\Theta = \mathrm{St}(n,m) = \{\theta \in \mathrm{Mat}(n,m) : \theta^\top \theta = I_{m\times m}\}$$

The group $S = \mathrm{SO}(n)$ preserves this manifold. And given a loss function $\ell : \Theta \to \mathbb{R}_{\geq 0}$

$Y \in \mathfrak{so}(n)$ the update direction $\qquad Y = \mathrm{Skew Y_0} = \frac{Y_0 - Y_0^\top}{2} \qquad Y_0 = \mathbb{E}_{q_\Lambda}[\nabla \ell \theta^\top]$

Here the distributions are parametrized by $\Lambda \in \mathrm{Mat}(n,m)$

$$q_\Lambda(\theta) \propto e^{-\operatorname{Tr}(\Lambda^\top \theta)}$$

and the update is given by

$\Lambda \leftarrow e^{-\alpha Y}\Lambda \qquad$ (actually an efficient variation is used)

# Koichi Tojo, Taro Yoshino's: "*Harmonic Exponential Families*".

$G$ a Lie group $H \leq G$. Let $\nu$ be a relatively invariant measure on $G$ $\pi : G \to \mathrm{GL}(V)$ a representation of $G$. Let $\alpha$ be a 1-cocycle of $\pi$ such that $\alpha\big|_H \equiv 0$. So $\alpha : G \to V$ satisfies

$$\alpha(gh) = \pi(g)\alpha(h) + \alpha(g) = \alpha(g). \qquad \text{So } \alpha : \underbrace{G/H}_{:=\Theta} \to V$$

Let $\nu$ be a relatively invariant measure on $\Theta$, meaning $\nu(gE) = \chi(g)\nu(E)$ for some homomorphism $\chi$. Let $\lambda \in V^\vee$ s.t. $A(\lambda) = \log \int_\Theta e^{-\langle \lambda, \alpha(\theta)\rangle} \mathrm{d}\nu(\theta) < \infty$. For such $\lambda$

$$q_\lambda(\theta)\mathrm{d}\nu(\theta) := e^{-\langle \lambda, \alpha(\theta)\rangle - A(\lambda)} \mathrm{d}\nu(\theta)$$

forms an exponential family satisfying

$$\frac{1}{\chi(g)} q_\lambda(g^{-1}\theta) := q_{\pi^\vee(g)\lambda}(\theta) \qquad \text{where } \langle \pi^\vee(g)\lambda, v\rangle = \langle \lambda, \pi(g)v\rangle.$$

# Constrained maximization: Statistical mechanics interpretation

Assume $\theta$ to be a kind of "microstate" with energy level $\ell(\theta)$. So $\Theta$ is some "state space".

# Constrained maximization: Statistical mechanics interpretation

Assume $\theta$ to be a kind of "microstate" with energy level $\ell(\theta)$. So $\Theta$ is some "state space".

Statistical mechanics: Assume a distribution of the microstates (across "particles") maximizing entropy, constrained to have expected energy $\leq E_0$.

# Constrained maximization: Statistical mechanics interpretation

Assume $\theta$ to be a kind of "microstate" with energy level $\ell(\theta)$. So $\Theta$ is some "state space".

Statistical mechanics: Assume a distribution of the microstates (across "particles") maximizing entropy, constrained to have expected energy $\leq E_0$.

Lagrange multiplier $\beta \geq 0$:

$$\arg\min_{q \in \mathcal{P}_\nu(\Theta)} -\mathcal{H}_\nu(q) + \beta(\mathbb{E}_{q\mathrm{d}\nu}[\ell] - E_0) = \arg\min_{q \in \mathcal{P}_\nu(\Theta)} \mathbb{E}_{q\mathrm{d}\nu}[\ell] - \frac{1}{\beta}\mathcal{H}_\nu(q)$$

$\tau = \frac{1}{\beta}$ corresponds to the thermodynamical notion of temperature.

## The exact posterior.

If $\mathcal{Q} = \mathcal{P}_\nu(\Theta)$ then there is a unique minimizer $p_\tau(\boldsymbol{\theta}) \propto e^{-\frac{1}{\tau}\ell(\boldsymbol{\theta})}$:

$$\arg\min_{q \in \mathcal{Q}} \mathbb{E}_{q\mathrm{d}\nu}[\ell] - \tau\mathcal{H}(q) =$$

# The exact posterior.

If $\mathcal{Q} = \mathcal{P}_\nu(\Theta)$ then there is a unique minimizer $p_\tau(\boldsymbol{\theta}) \propto e^{-\frac{1}{\tau}\ell(\boldsymbol{\theta})}$:

$$\arg\min_{q \in \mathcal{Q}} \mathbb{E}_{q\mathrm{d}\nu}[\ell] - \tau\mathcal{H}(q) = \arg\min_{q \in \mathcal{Q}} \mathbb{E}_{q\mathrm{d}\nu}\left[-\log e^{-\frac{1}{\tau}\ell}\right] + \mathbb{E}_{q\mathrm{d}\nu}[\log q]$$

# The exact posterior.

If $\mathcal{Q} = \mathcal{P}_\nu(\Theta)$ then there is a unique minimizer $p_\tau(\boldsymbol{\theta}) \propto e^{-\frac{1}{\tau}\ell(\boldsymbol{\theta})}$:

$$\arg\min_{q \in \mathcal{Q}} \mathbb{E}_{q\mathrm{d}\nu}[\ell] - \tau \mathcal{H}(q) = \arg\min_{q \in \mathcal{Q}} \int_\Theta \log\left(\frac{q(\boldsymbol{\theta})}{e^{-\frac{1}{\tau}\ell(\boldsymbol{\theta})}}\right) q(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta})$$

# The exact posterior.

If $\mathcal{Q} = \mathcal{P}_\nu(\Theta)$ then there is a unique minimizer $p_\tau(\boldsymbol{\theta}) \propto e^{-\frac{1}{\tau}\ell(\boldsymbol{\theta})}$:

$$\arg\min_{q \in \mathcal{Q}} \mathbb{E}_{q\mathrm{d}\nu}[\ell] - \tau\mathcal{H}(q) = \arg\min_{q \in \mathcal{Q}} \int_\Theta \log\left(\frac{q(\boldsymbol{\theta})}{p_\tau(\boldsymbol{\theta})}\right) q(\boldsymbol{\theta})\mathrm{d}\nu(\boldsymbol{\theta}) + \text{const.}$$

# The exact posterior.

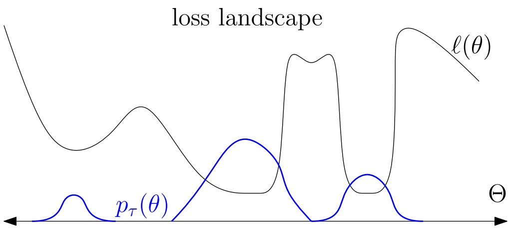
If $\mathcal{Q} = \mathcal{P}_\nu(\Theta)$ then there is a unique minimizer $p_\tau(\boldsymbol{\theta}) \propto e^{-\frac{1}{\tau}\ell(\boldsymbol{\theta})}$:

$$\arg\min_{q \in \mathcal{Q}} \mathbb{E}_{q\mathrm{d}\nu}[\ell] - \tau \mathcal{H}(q) = \arg\min_{q \in \mathcal{Q}} \mathbb{D}_\nu(q \| p_\tau).$$

# The exact posterior.

If $\mathcal{Q} = \mathcal{P}_\nu(\Theta)$ then there is a unique minimizer $p_\tau(\boldsymbol{\theta}) \propto e^{-\frac{1}{\tau}\ell(\boldsymbol{\theta})}$:

$$\arg\min_{q\in\mathcal{Q}} \mathbb{E}_{q\mathrm{d}\nu}[\ell] - \tau\mathcal{H}(q) = \arg\min_{q\in\mathcal{Q}} \mathbb{D}_\nu(q\|p_\tau).$$

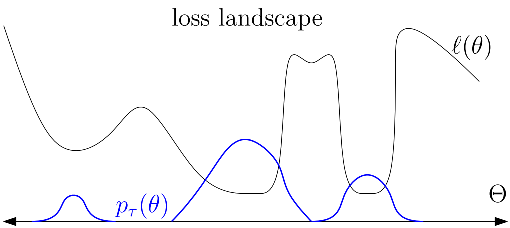


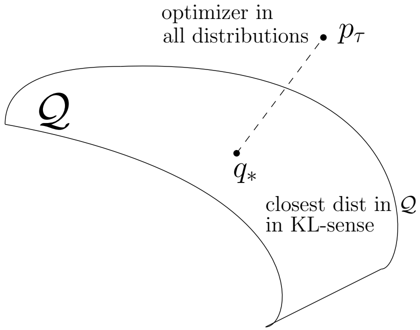

Minimize the objective $\mathcal{E}(q) := \mathbb{D}(q\|p_\tau)$ for $q \in \mathcal{Q}$...

# The exact posterior.

If $\mathcal{Q} = \mathcal{P}_\nu(\Theta)$ then there is a unique minimizer $p_\tau(\boldsymbol{\theta}) \propto e^{-\frac{1}{\tau}\ell(\boldsymbol{\theta})}$:

$$\arg\min_{q\in\mathcal{Q}} \mathbb{E}_{q\mathrm{d}\nu}[\ell] - \tau\mathcal{H}(q) = \arg\min_{q\in\mathcal{Q}} \mathbb{D}_\nu(q\|p_\tau).$$




Minimize the objective $\mathcal{E}(q) := \mathbb{D}(q\|p_\tau)$ for $q \in \mathcal{Q}$...

...an approximate Bayesian solution.

# Why call it Bayesian?

Let $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}) + R(\boldsymbol{\theta})$. Observe new data $(\mathbf{x}_{\text{new}}, y_{\text{new}})$ with loss contribution $\ell_{\text{new}}$.

# Why call it Bayesian?

Let $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}) + R(\boldsymbol{\theta})$. Observe new data $(\mathbf{x}_{\text{new}}, y_{\text{new}})$ with loss contribution $\ell_{\text{new}}$.

How to update $p_\tau$? Take $\tau = 1$

# Why call it Bayesian?

Let $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}) + R(\boldsymbol{\theta})$. Observe new data $(\mathbf{x}_{\text{new}}, y_{\text{new}})$ with loss contribution $\ell_{\text{new}}$.

How to update $p_\tau$? Take $\tau = 1$

Bayes' rule is about conditional probabilities, and updating priors:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Interpret $e^{-\ell_i(\boldsymbol{\theta})}$ as the likelihood of observing label $y_i$ given the model parameter $\boldsymbol{\theta}$ and $\mathbf{x}_i$.
Interpret $\pi(\boldsymbol{\theta}) \propto e^{-R(\boldsymbol{\theta})}$ as the prior on the parameters.

# Why call it Bayesian?

Let $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}) + R(\boldsymbol{\theta})$. Observe new data $(\mathbf{x}_{\text{new}}, y_{\text{new}})$ with loss contribution $\ell_{\text{new}}$.

How to update $p_\tau$? Take $\tau = 1$

Bayes' rule is about conditional probabilities, and updating priors:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Interpret $e^{-\ell_i(\boldsymbol{\theta})}$ as the likelihood of observing label $y_i$ given the model parameter $\boldsymbol{\theta}$ and $\mathbf{x}_i$.
Interpret $\pi(\boldsymbol{\theta}) \propto e^{-R(\boldsymbol{\theta})}$ as the prior on the parameters.

After one round of learning the posterior $p \propto e^{-\sum_i \ell_i}\pi$ is our prior belief about $\boldsymbol{\theta}$ distribution. According to Bayes rule updated belief should be *after* a new data point.

$$p_{\text{updated}}(\boldsymbol{\theta}) \propto e^{-\ell_{\text{new}}(\boldsymbol{\theta})} p(\boldsymbol{\theta}).$$

# Why call it Bayesian?

Let $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}) + R(\boldsymbol{\theta})$. Observe new data $(\mathbf{x}_{\text{new}}, y_{\text{new}})$ with loss contribution $\ell_{\text{new}}$.

How to update $p_\tau$? Take $\tau = 1$

Bayes' rule is about conditional probabilities, and updating priors:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Interpret $e^{-\ell_i(\boldsymbol{\theta})}$ as the likelihood of observing label $y_i$ given the model parameter $\boldsymbol{\theta}$ and $\mathbf{x}_i$.
Interpret $\pi(\boldsymbol{\theta}) \propto e^{-R(\boldsymbol{\theta})}$ as the prior on the parameters.

After one round of learning the posterior $p \propto e^{-\sum_i \ell_i}\pi$ is our prior belief about $\boldsymbol{\theta}$ distribution.
According to Bayes rule updated belief should be *after* a new data point.

$$p_{\text{updated}}(\boldsymbol{\theta}) \propto e^{-\ell_{\text{new}}(\boldsymbol{\theta})} p(\boldsymbol{\theta}).$$

This is also the optimizer if we had initially considered the loss function $\ell_{\text{updated}} = \ell + \ell_{\text{new}}$.

# Exponential Families

Let $T : \Theta \to V$, called the sufficient statistic. Call

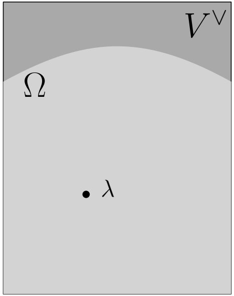
$$\Omega = \Omega_\nu(T) = \left\{ \lambda \in V^\vee : A(\lambda) := \log \int_\Theta e^{-\langle \lambda, T(\boldsymbol{\theta}) \rangle} \mathrm{d}\nu(\boldsymbol{\theta}) < \infty \right\}.$$

Then $q_\lambda(\boldsymbol{\theta}) = e^{-\langle \lambda, T(\boldsymbol{\theta}) \rangle - A(\lambda)}$ form an exponential family of distributions.

# Exponential Families

Let $T : \Theta \to V$, called the sufficient statistic. Call

$$\Omega = \Omega_\nu(T) = \left\{ \lambda \in V^\vee : A(\lambda) := \log \int_\Theta e^{-\langle \lambda, T(\boldsymbol{\theta}) \rangle} \mathrm{d}\nu(\boldsymbol{\theta}) < \infty \right\}.$$




Then $q_\lambda(\boldsymbol{\theta}) = e^{-\langle \lambda, T(\boldsymbol{\theta})\rangle - A(\lambda)}$ form an exponential family of distributions.

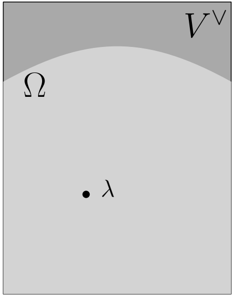
$$-\frac{\partial A}{\partial \lambda_i} = \int_\Theta T_i(\boldsymbol{\theta}) e^{-\langle \lambda, T(\boldsymbol{\theta})\rangle - A(\lambda)} \mathrm{d}\nu(\boldsymbol{\theta}) = \mathbb{E}_{q_\lambda \mathrm{d}\nu}[T_i] =: \mu_i$$

$$\frac{\partial^2 A}{\partial \lambda_i \partial \lambda_j} = \int_\Theta (T_i(\boldsymbol{\theta}) - \mu_i)(T_j(\boldsymbol{\theta}) - \mu_j) q_\lambda(\boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{\theta})$$

$$= \mathbb{E}_{q_\lambda} \left[ \left( \tfrac{\partial}{\partial \lambda_i} \log q_\lambda \right) \left( \tfrac{\partial}{\partial \lambda_j} \log q_\lambda \right) \right] =: F_{i,j}(\lambda) \qquad \text{Fisher Matrix}$$

Let $T : \Theta \to V$, called the sufficient statistic. Call

$$\Omega = \Omega_\nu(T) = \left\{ \lambda \in V^\vee : A(\lambda) := \log \int_\Theta e^{-\langle \lambda, T(\boldsymbol{\theta}) \rangle} \mathrm{d}\nu(\boldsymbol{\theta}) < \infty \right\}.$$




Then $q_\lambda(\boldsymbol{\theta}) = e^{-\langle \lambda, T(\boldsymbol{\theta}) \rangle - A(\lambda)}$ form an exponential family of distributions.

$$\begin{aligned}
-\frac{\partial A}{\partial \lambda_i} &= \int_\Theta T_i(\boldsymbol{\theta}) e^{-\langle \lambda, T(\boldsymbol{\theta}) \rangle - A(\lambda)} \mathrm{d}\nu(\boldsymbol{\theta}) = \mathbb{E}_{q_\lambda \mathrm{d}\nu}[T_i] =: \mu_i \\
\frac{\partial^2 A}{\partial \lambda_i \partial \lambda_j} &= \int_\Theta (T_i(\boldsymbol{\theta}) - \mu_i)(T_j(\boldsymbol{\theta}) - \mu_j) q_\lambda(\boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{\theta}) \\
&= \mathbb{E}_{q_\lambda}\left[ \left( \tfrac{\partial}{\partial \lambda_i} \log q_\lambda \right) \left( \tfrac{\partial}{\partial \lambda_j} \log q_\lambda \right) \right] =: F_{i,j}(\lambda) \qquad \text{Fisher Matrix}
\end{aligned}$$

Example: If $T(\boldsymbol{\theta}) = \left[ \begin{smallmatrix} \boldsymbol{\theta} \\ \boldsymbol{\theta}^2 \end{smallmatrix} \right]$ then we get 1-D Gaussians $q_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \propto e^{-\lambda_1 \boldsymbol{\theta} - \lambda_2 \boldsymbol{\theta}^2}$ for $\lambda_2 > 0$.

# Harmonic exponential families (by Koichi Tojo & Taro Yoshino)

What about exponential families on $\Theta$ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.

# Harmonic exponential families (by Koichi Tojo & Taro Yoshino)

What about exponential families on $\Theta$ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.
- $\nu$ a relatively invariant base measure
  $\mathrm{d}\nu(g \cdot \theta) = \chi(g)\mathrm{d}\nu(\theta)$.

# Harmonic exponential families (by Koichi Tojo & Taro Yoshino)

What about exponential families on $\Theta$ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.
- $\nu$ a relatively invariant base measure
  $\mathrm{d}\nu(g \cdot \theta) = \chi(g)\mathrm{d}\nu(\theta)$.
- A finite dimensional representation
  $\pi : G \to \mathrm{GL}(V)$.

# Harmonic exponential families (by Koichi Tojo & Taro Yoshino)

What about exponential families on $\Theta$ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.
- $\nu$ a relatively invariant base measure $\mathrm{d}\nu(g \cdot \theta) = \chi(g)\mathrm{d}\nu(\theta)$.
- A finite dimensional representation $\pi : G \to \mathrm{GL}(V)$.
- A 1-cocycle of $\pi$ such that $\alpha\big|_H \equiv 0$. So $\alpha : G \to V$ satisfies

$$\alpha(gh) = \pi(g)\alpha(h) + \alpha(g) = \alpha(g).$$

Thus $\alpha : \Theta \to V$.

# Harmonic exponential families (by Koichi Tojo & Taro Yoshino)

What about exponential families on $\Theta$ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.
- $\nu$ a relatively invariant base measure $\mathrm{d}\nu(g \cdot \theta) = \chi(g)\mathrm{d}\nu(\theta)$.
- A finite dimensional representation $\pi : G \to \mathrm{GL}(V)$.
- A 1-cocycle of $\pi$ such that $\alpha\big|_H \equiv 0$. So $\alpha : G \to V$ satisfies

$$\alpha(gh) = \pi(g)\alpha(h) + \alpha(g) = \alpha(g).$$

Thus $\alpha : \Theta \to V$.

Let $\lambda \in \Omega_\nu(\alpha) \subseteq V^\vee$ i.e.,
$A(\lambda) = \log \int_\Theta e^{-\langle \lambda, \alpha(\theta)\rangle} \mathrm{d}\nu(\theta) < \infty$.

$$q_\lambda(\theta)\mathrm{d}\nu(\theta) := e^{-\langle \lambda, \alpha(\theta)\rangle - A(\lambda)} \mathrm{d}\nu(\theta)$$

# Harmonic exponential families (by Koichi Tojo & Taro Yoshino)

What about exponential families on $\Theta$ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.
- $\nu$ a relatively invariant base measure $\mathrm{d}\nu(g \cdot \theta) = \chi(g)\mathrm{d}\nu(\theta)$.
- A finite dimensional representation $\pi : G \to \mathrm{GL}(V)$.
- A 1-cocycle of $\pi$ such that $\alpha\big|_H \equiv 0$. So $\alpha : G \to V$ satisfies

$$\alpha(gh) = \pi(g)\alpha(h) + \alpha(g) = \alpha(g).$$

Thus $\alpha : \Theta \to V$.

Let $\lambda \in \Omega_\nu(\alpha) \subseteq V^\vee$ i.e.,
$A(\lambda) = \log \int_\Theta e^{-\langle \lambda, \alpha(\theta) \rangle} \mathrm{d}\nu(\theta) < \infty$.

$$q_\lambda(\theta)\mathrm{d}\nu(\theta) := e^{-\langle \lambda, \alpha(\theta) \rangle - A(\lambda)} \mathrm{d}\nu(\theta)$$

forms an exponential family closed under pushforwards

# Harmonic exponential families (by Koichi Tojo & Taro Yoshino)

What about exponential families on $\Theta$ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.
- $\nu$ a relatively invariant base measure $\mathrm{d}\nu(g \cdot \theta) = \chi(g)\mathrm{d}\nu(\theta)$.
- A finite dimensional representation $\pi : G \to \mathrm{GL}(V)$.
- A 1-cocycle of $\pi$ such that $\alpha\big|_H \equiv 0$. So $\alpha : G \to V$ satisfies

$$\alpha(gh) = \pi(g)\alpha(h) + \alpha(g) = \alpha(g).$$

Thus $\alpha : \Theta \to V$.

Let $\lambda \in \Omega_\nu(\alpha) \subseteq V^\vee$ i.e.,
$A(\lambda) = \log \int_\Theta e^{-\langle \lambda, \alpha(\theta) \rangle} \mathrm{d}\nu(\theta) < \infty$.

$$q_\lambda(\theta)\mathrm{d}\nu(\theta) := e^{-\langle \lambda, \alpha(\theta) \rangle - A(\lambda)} \mathrm{d}\nu(\theta)$$

forms an exponential family closed under pushforwards

$$(q_\lambda)^g(\theta) = \frac{1}{\chi(g)} q_\lambda(g^{-1}\theta)$$

# Harmonic exponential families (by Koichi Tojo & Taro Yoshino)

What about exponential families on $\Theta$ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.
- $\nu$ a relatively invariant base measure $\mathrm{d}\nu(g\cdot\theta) = \chi(g)\mathrm{d}\nu(\theta)$.
- A finite dimensional representation $\pi : G \to \mathrm{GL}(V)$.
- A 1-cocycle of $\pi$ such that $\alpha\big|_H \equiv 0$. So $\alpha : G \to V$ satisfies

$$\alpha(gh) = \pi(g)\alpha(h) + \alpha(g) = \alpha(g).$$

Thus $\alpha : \Theta \to V$.

Let $\lambda \in \Omega_\nu(\alpha) \subseteq V^\vee$ i.e.,
$A(\lambda) = \log \int_\Theta e^{-\langle\lambda,\alpha(\theta)\rangle}\mathrm{d}\nu(\theta) < \infty$.

$$q_\lambda(\theta)\mathrm{d}\nu(\theta) := e^{-\langle\lambda,\alpha(\theta)\rangle - A(\lambda)}\mathrm{d}\nu(\theta)$$

forms an exponential family closed under pushforwards

$$(q_\lambda)^g(\theta) = \frac{1}{\chi(g)}q_\lambda(g^{-1}\theta) = e^{-\langle\lambda,\alpha(g^{-1}\theta)\rangle - A(\lambda)}$$

# Harmonic exponential families (by Koichi Tojo & Taro Yoshino)

What about exponential families on $\Theta$ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.
- $\nu$ a relatively invariant base measure $\mathrm{d}\nu(g \cdot \theta) = \chi(g)\mathrm{d}\nu(\theta)$.
- A finite dimensional representation $\pi : G \to \mathrm{GL}(V)$.
- A 1-cocycle of $\pi$ such that $\alpha\big|_H \equiv 0$. So $\alpha : G \to V$ satisfies

$$\alpha(gh) = \pi(g)\alpha(h) + \alpha(g) = \alpha(g).$$

Thus $\alpha : \Theta \to V$.

Let $\lambda \in \Omega_\nu(\alpha) \subseteq V^\vee$ i.e.,
$A(\lambda) = \log \int_\Theta e^{-\langle \lambda, \alpha(\theta) \rangle} \mathrm{d}\nu(\theta) < \infty$.

$$q_\lambda(\theta)\mathrm{d}\nu(\theta) := e^{-\langle \lambda, \alpha(\theta)\rangle - A(\lambda)}\mathrm{d}\nu(\theta)$$

forms an exponential family closed under pushforwards

$$(q_\lambda)^g(\theta) = \frac{1}{\chi(g)} q_\lambda(g^{-1}\theta) = e^{-\langle \lambda, \alpha(g^{-1}\theta)\rangle - A(\lambda)}$$
$$= e^{-\langle \lambda, \pi(g^{-1})\alpha(\theta)\rangle - A(\lambda) - \alpha(g^{-1})}$$

# Harmonic exponential families (by Koichi Tojo & Taro Yoshino)

What about exponential families on $\Theta$ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.
- $\nu$ a relatively invariant base measure $\mathrm{d}\nu(g \cdot \theta) = \chi(g)\mathrm{d}\nu(\theta)$.
- A finite dimensional representation $\pi : G \to \mathrm{GL}(V)$.
- A $1$-cocycle of $\pi$ such that $\alpha\big|_H \equiv 0$. So $\alpha : G \to V$ satisfies

$$\alpha(gh) = \pi(g)\alpha(h) + \alpha(g) = \alpha(g).$$

Thus $\alpha : \Theta \to V$.

Let $\lambda \in \Omega_\nu(\alpha) \subseteq V^\vee$ i.e., $A(\lambda) = \log \int_\Theta e^{-\langle \lambda, \alpha(\theta) \rangle} \mathrm{d}\nu(\theta) < \infty$.

$$q_\lambda(\theta)\mathrm{d}\nu(\theta) := e^{-\langle \lambda, \alpha(\theta) \rangle - A(\lambda)} \mathrm{d}\nu(\theta)$$

forms an exponential family closed under pushforwards

$$\begin{aligned}
(q_\lambda)^g(\theta) &= \frac{1}{\chi(g)} q_\lambda(g^{-1}\theta) = e^{-\langle \lambda, \alpha(g^{-1}\theta) \rangle - A(\lambda)} \\
&= e^{-\langle \lambda, \pi(g^{-1})\alpha(\theta) \rangle - A(\lambda) - \alpha(g^{-1})} \\
&\propto e^{-\langle \pi^\vee(g)\lambda, \alpha(\theta) \rangle}
\end{aligned}$$

# Harmonic exponential families (by Koichi Tojo & Taro Yoshino)

What about exponential families on $\Theta$ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.
- $\nu$ a relatively invariant base measure $\mathrm{d}\nu(g \cdot \theta) = \chi(g)\mathrm{d}\nu(\theta)$.
- A finite dimensional representation $\pi : G \to \mathrm{GL}(V)$.
- A 1-cocycle of $\pi$ such that $\alpha\big|_H \equiv 0$. So $\alpha : G \to V$ satisfies

$$\alpha(gh) = \pi(g)\alpha(h) + \alpha(g) = \alpha(g).$$

Thus $\alpha : \Theta \to V$.

Let $\lambda \in \Omega_\nu(\alpha) \subseteq V^\vee$ i.e., $A(\lambda) = \log \int_\Theta e^{-\langle \lambda, \alpha(\theta) \rangle} \mathrm{d}\nu(\theta) < \infty$.

$$q_\lambda(\theta)\mathrm{d}\nu(\theta) := e^{-\langle \lambda, \alpha(\theta)\rangle - A(\lambda)}\mathrm{d}\nu(\theta)$$

forms an exponential family closed under pushforwards

$$\begin{aligned}(q_\lambda)^g(\theta) &= \frac{1}{\chi(g)} q_\lambda(g^{-1}\theta) = e^{-\langle \lambda, \alpha(g^{-1}\theta)\rangle - A(\lambda)} \\ &= e^{-\langle \lambda, \pi(g^{-1})\alpha(\theta)\rangle - A(\lambda) - \alpha(g^{-1})} \\ &\propto e^{-\langle \pi^\vee(g)\lambda, \alpha(\theta)\rangle} \propto q_{\pi^\vee(g)\lambda}(\theta)\end{aligned}$$

# Linear Approximation of Lie group update is NGD on natural parameters

The Lie group rule is for transformation families. BLR is NGD on $\lambda$'s of exponential families.

# Linear Approximation of Lie group update is NGD on natural parameters

The Lie group rule is for transformation families. BLR is NGD on $\lambda$'s of exponential families. The overlap is harmonic exponential families:

Pushforwards of $q_\lambda$ are still in the family,

$$(q_\lambda)^g = q_{\lambda'} \quad \text{with} \quad \lambda' = \pi^\vee(g)\lambda$$

# Linear Approximation of Lie group update is NGD on natural parameters

The Lie group rule is for transformation families. BLR is NGD on $\lambda$'s of exponential families. The overlap is harmonic exponential families:

Pushforwards of $q_\lambda$ are still in the family,

$$(q_\lambda)^g = q_{\lambda'} \quad \text{with} \quad \lambda' = \pi^\vee(g)\lambda$$

Thus $\widetilde{Q} = \{\lambda \in \Omega : \lambda = \pi^\vee(g)\lambda_0, g \in G\}$ and updates are given by $\lambda^{\text{updated}} = \pi^\vee(g^{\text{updated}})\lambda_0$.

# Linear Approximation of Lie group update is NGD on natural parameters

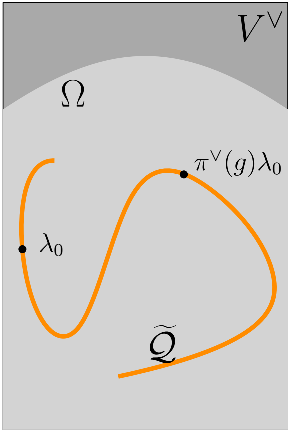
The Lie group rule is for transformation families. BLR is NGD on $\lambda$'s of exponential families. The overlap is harmonic exponential families:

Pushforwards of $q_\lambda$ are still in the family,

$$(q_\lambda)^g = q_{\lambda'} \quad \text{with} \quad \lambda' = \pi^\vee(g)\lambda$$

Thus $\widetilde{Q} = \{\lambda \in \Omega : \lambda = \pi^\vee(g)\lambda_0, g \in G\}$ and updates are given by $\lambda^{\text{updated}} = \pi^\vee(g^{\text{updated}})\lambda_0$.
Other quantities of $q_\lambda$ also vary with $g$:

$$\mu(\lambda') = \pi(g)\lambda + \alpha(g)$$
$$A(\lambda') = A(\lambda) + \log(\chi(g)) + \alpha(g^{-1})$$



# Linear Approximation of Lie group update is NGD on natural parameters

The Lie group rule is for transformation families. BLR is NGD on $\lambda$'s of exponential families. The overlap is harmonic exponential families:

Pushforwards of $q_\lambda$ are still in the family,

$$(q_\lambda)^g = q_{\lambda'} \quad \text{with} \quad \lambda' = \pi^\vee(g)\lambda$$

Thus $\widetilde{Q} = \{\lambda \in \Omega : \lambda = \pi^\vee(g)\lambda_0, g \in G\}$ and updates are given by $\lambda^{\text{updated}} = \pi^\vee(g^{\text{updated}})\lambda_0$. Other quantities of $q_\lambda$ also vary with $g$:

$$\mu(\lambda') = \pi(g)\lambda + \alpha(g)$$
$$A(\lambda') = A(\lambda) + \log(\chi(g)) + \alpha(g^{-1})$$