Lie Group Bayesian Learning

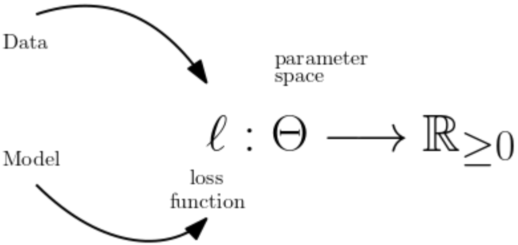
Work of: E. Mehmet Kıral[1], and Thomas Möllenhoff[2], M. Emtiyaz Khan[2]
Keigo Nishida[2], Koichi Tojo[1], Kenichi Bannai[1].

September 3, 2024 CALISTA workshop
Geometry Involved Machine Learning

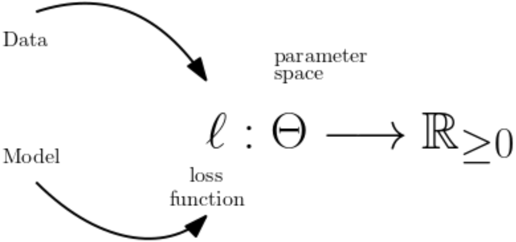
The work is supported mainly by the Bayes-duality project, JST CREST Grant Number JPMJCR2112.

[1] Keio University, [2] RIKEN AIP

The classical and Bayesian learning setups



The classical and Bayesian learning setups



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

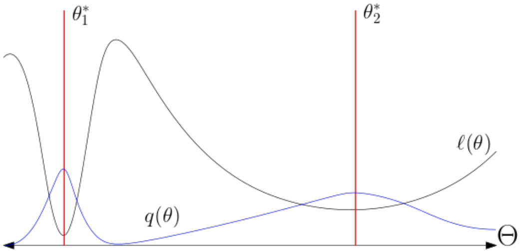
Bayesian : find a distribution $q \in \mathcal{P}(\Theta)$ ....

Classical vs. Bayesian learning

The loss function is highly nonconvex. Usually

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

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



θ_1^* and θ_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

$\ell(\theta)$, a loss function on model parameters $\theta \in \Theta$. Pick a base measure ν on Θ.

The Bayesian Learning Problem

$\ell(\theta)$, a loss function on model parameters $\theta \in \Theta$. Pick a base measure ν on Θ. We solve

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

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

The Bayesian Learning Problem

$\ell(\theta)$, a loss function on model parameters $\theta \in \Theta$. Pick a base measure ν on Θ. We solve

$$q_* \in \underset{q \in \mathcal{Q}}{\arg\min} \; \mathbb{E}_q[\ell] - \tau \mathcal{H}_\nu(q)$$

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

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

The Bayesian Learning Problem

$\ell(\theta)$, a loss function on model parameters $\theta \in \Theta$. Pick a base measure ν on Θ. We solve

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

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

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

The Bayesian Learning Problem

$\ell(\theta)$, a loss function on model parameters $\theta \in \Theta$. Pick a base measure ν on Θ. We solve

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

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

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

The exact posterior.

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

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

The exact posterior.

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

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

The exact posterior.

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

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

The exact posterior.

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

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

The exact posterior.

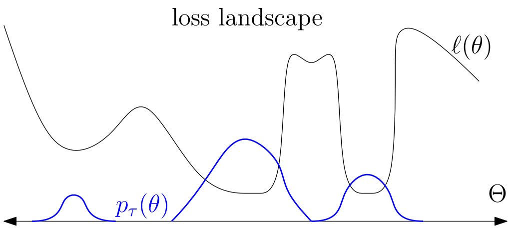
If $\mathcal{Q} = \mathcal{P}_\nu(\Theta)$ then there is a unique minimizer $p_\tau(\theta) \propto e^{-\frac{1}{\tau}\ell(\theta)}$:

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

The exact posterior.

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

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

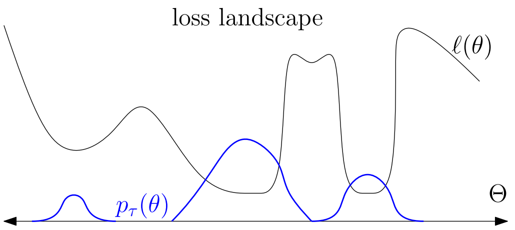


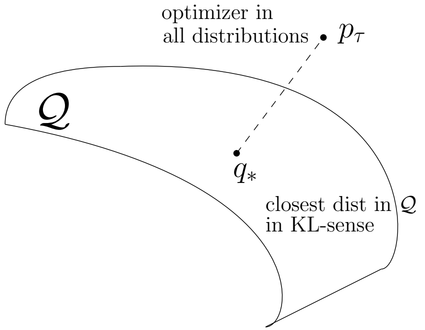
Minimize the objective $\mathcal{E}(q) := \mathbb{D}(q \| p_\tau)$ for $q \in \mathcal{Q}$...

The exact posterior.

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

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





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

...an approximate Bayesian solution.

Previously... BLR: The Bayesian Learning Rule, [KR21][1]

[KR21] take $\mathcal{Q}$ as exponential families are $q_\lambda(\theta) \propto e^{-\lambda^\top T(\theta)}$.
$T : \Theta \to \mathbb{R}^d$ is a *sufficient statistic*, λ are *natural parameters*.

Gaussians, Exponential distributions, Gamma, inverse Gamma, Wishart, von-Mises, etc.

[1] [KR21]: Khan, M. E. and Rue H., *The Bayesian Learning Rule*

Previously...BLR: The Bayesian Learning Rule, [KR21][1]

[KR21] take $\mathcal{Q}$ as exponential families are $q_\lambda(\theta) \propto e^{-\lambda^\top T(\theta)}$.
$T : \Theta \to \mathbb{R}^d$ is a *sufficient statistic*, λ are *natural parameters*.

$$\lambda \longleftarrow \lambda - \alpha F(\lambda)^{-1} \nabla_\lambda \mathcal{E}(q_\lambda)$$

BLR is Natural Gradient Descent on λ parameters.

Gaussians, Exponential distributions, Gamma, inverse Gamma, Wishart, von-Mises, etc.

$\alpha > 0$ step size
$F(\lambda)$ the Fisher matrix

[1] [KR21]: Khan, M. E. and Rue H., *The Bayesian Learning Rule*

Gaussians, Exponential distributions, Gamma, inverse Gamma, Wishart, von-Mises, etc.

[KR21] take $\mathcal{Q}$ as exponential families are $q_\lambda(\theta) \propto e^{-\lambda^\top T(\theta)}$.
$T : \Theta \to \mathbb{R}^d$ is a *sufficient statistic*, λ are *natural parameters*.

$$\lambda \longleftarrow \lambda - \alpha F(\lambda)^{-1} \nabla_\lambda \mathcal{E}(q_\lambda)$$

$\alpha > 0$ step size
$F(\lambda)$ the Fisher matrix

BLR is Natural Gradient Descent on λ parameters.

Issue 1 The candidates $\mathcal{Q}$ is required to be an exponential family,

[1] [KR21]: Khan, M. E. and Rue H., *The Bayesian Learning Rule*

Previously...BLR: The Bayesian Learning Rule, [KR21][1]

[KR21] take $\mathcal{Q}$ as exponential families are $q_\lambda(\theta) \propto e^{-\lambda^\top T(\theta)}$.
$T : \Theta \to \mathbb{R}^d$ is a *sufficient statistic*, λ are *natural parameters*.

Gaussians, Exponential distributions, Gamma, inverse Gamma, Wishart, von-Mises, etc.

$$\lambda \longleftarrow \lambda - \alpha F(\lambda)^{-1} \nabla_\lambda \mathcal{E}(q_\lambda)$$

$\alpha > 0$ step size
$F(\lambda)$ the Fisher matrix

BLR is Natural Gradient Descent on λ parameters.

- **Issue 1** The candidates $\mathcal{Q}$ is required to be an exponential family,
- **Issue 2** Not every λ is allowed as a natural parameter, and the linear update rule could overshoot the constraints.

[1] [KR21]: Khan, M. E. and Rue H., *The Bayesian Learning Rule*

Previously... BLR: The Bayesian Learning Rule, [KR21][1]

[KR21] take $\mathcal{Q}$ as exponential families are $q_\lambda(\theta) \propto e^{-\lambda^\top T(\theta)}$.
$T : \Theta \to \mathbb{R}^d$ is a *sufficient statistic*, λ are *natural parameters*.

Gaussians, Exponential distributions, Gamma, inverse Gamma, Wishart, von-Mises, etc.

$$\lambda \longleftarrow \lambda - \alpha F(\lambda)^{-1} \nabla_\lambda \mathcal{E}(q_\lambda)$$

$\alpha > 0$ step size
$F(\lambda)$ the Fisher matrix

BLR is Natural Gradient Descent on λ parameters.

- Issue 1 The candidates $\mathcal{Q}$ is required to be an exponential family,
- Issue 2 Not every λ is allowed as a natural parameter, and the linear update rule could overshoot the constraints.
- Issue 3 Computing $\nabla_\lambda \mathcal{E}(q_\lambda)$ is not efficient in general but for special exponential families.

[1][KR21]: Khan, M. E. and Rue H., *The Bayesian Learning Rule*

Parametrizing Q by groups.

Assuma a Lie group G acts on the parameter manifold Θ,

Parametrizing Q by groups.

Assuma a Lie group G acts on the parameter manifold Θ, it also acts on distributions on Θ.

Parametrizing $\mathcal{Q}$ by groups.

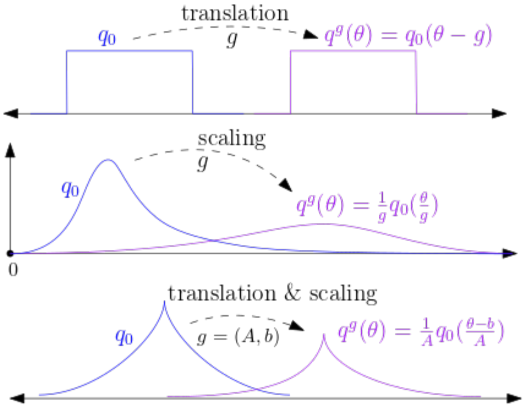
Assuma a Lie group G acts on the parameter manifold Θ, it also acts on distributions on Θ. $\mathcal{Q}$ is formed as the orbit of such an action for any base distribution q_0:

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

Parametrizing $\mathcal{Q}$ by groups.

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

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



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

Optimization on the group

We now solve

$$\arg\min_{g\in G} \mathcal{E}(q^g) = \arg\min_{g\in G} \int_\Theta q^g \log\left(\frac{q^g}{e^{-\frac{1}{\tau}\ell}}\right)$$

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

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(q^{ge^{tX}})\Big|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t} \underbrace{\int_\Theta q^{ge^{tX}}(\theta)\tfrac{1}{\tau}\ell(\theta)\mathrm{d}\nu(\theta)}_{\text{data contribution}} + \underbrace{\int_\Theta q^{ge^{tX}}(\theta)\log q^{ge^{tX}}(\theta)\mathrm{d}\nu(\theta)}_{\text{entropy contribution}}\Bigg|_{t=0}$$

Optimization on the group

We now solve

$$\arg\min_{g \in G} \mathcal{E}(q^g) = \arg\min_{g \in G} \int_\Theta q^g \log\left(\frac{q^g}{e^{-\frac{1}{\tau}\ell}}\right)$$

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

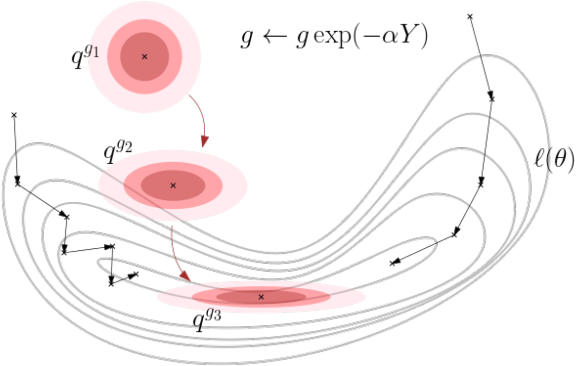
$$\frac{\mathrm{d}}{\mathrm{dt}} \mathcal{E}(q^{ge^{tX}})\Big|_{t=0} = \frac{\mathrm{d}}{\mathrm{dt}} \underbrace{\int_\Theta q^{ge^{tX}}(\theta)\frac{1}{\tau}\ell(\theta)\mathrm{d}\nu(\theta)}_{\text{data contribution}} + \underbrace{\int_\Theta q^{ge^{tX}}(\theta)\log q^{ge^{tX}}(\theta)\mathrm{d}\nu(\theta)}_{\text{entropy contribution}}\Bigg|_{t=0}$$

The data contribution can be rewritten as

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

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 θ.



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

Solved issues

[2] Mohamed et. al. *Monte carlo Gradient Estimation in Machine Learning* JMLR 2020

Solved issues

Issue 1 $\mathcal{Q}$ is required to be an exponential family.

Solution Can choose q_0 freely and push it around with a group.

[2] Mohamed et. al. *Monte carlo Gradient Estimation in Machine Learning* JMLR 2020

Solved issues

Issue 1 $\mathcal{Q}$ is required to be an exponential family.

Solution Can choose q_0 freely and push it around with a group.

Issue 2 The updates could overshoot and leave the manifold.

Solution Closure of the group under operation keeps updates on $\mathcal{Q}$.

[2] Mohamed et. al. *Monte carlo Gradient Estimation in Machine Learning* JMLR 2020

Solved issues

Issue 1 $\mathcal{Q}$ is required to be an exponential family.

Solution Can choose q_0 freely and push it around with a group.

Issue 2 The updates could overshoot and leave the manifold.

Solution Closure of the group under operation keeps updates on $\mathcal{Q}$.

Issue 3 The gradient $\nabla_\lambda \mathcal{E}(q_\lambda)$ can only be computed in special cases.

Solution Group action is the correct generality for reparametrization $\frac{\mathrm{d}}{\mathrm{d}g}\mathbb{E}_{q^g}[\ell] = \frac{\mathrm{d}}{\mathrm{d}g}\int_\Theta q^g(\theta)\ell(\theta)\mathrm{d}\nu(\theta)$.

[2] Mohamed et. al. *Monte carlo Gradient Estimation in Machine Learning* JMLR 2020

Solved issues

Issue 1 $\mathcal{Q}$ is required to be an exponential family.

Solution Can choose q_0 freely and push it around with a group.

Issue 2 The updates could overshoot and leave the manifold.

Solution Closure of the group under operation keeps updates on $\mathcal{Q}$.

Issue 3 The gradient $\nabla_\lambda \mathcal{E}(q_\lambda)$ can only be computed in special cases.

Solution Group action is the correct generality for reparametrization $\frac{\mathrm{d}}{\mathrm{d}g}\mathbb{E}_{q^g}[\ell] = \frac{\mathrm{d}}{\mathrm{d}g}\int_\Theta q_0(\theta)\ell(g\cdot\theta)\mathrm{d}\nu(\theta)$ (by changing $\theta \mapsto g\cdot\theta$)

[2] Mohamed et. al. *Monte carlo Gradient Estimation in Machine Learning* JMLR 2020

Solved issues

Issue 1 $\mathcal{Q}$ is required to be an exponential family.

Solution Can choose q_0 freely and push it around with a group.

Issue 2 The updates could overshoot and leave the manifold.

Solution Closure of the group under operation keeps updates on $\mathcal{Q}$.

Issue 3 The gradient $\nabla_\lambda \mathcal{E}(q_\lambda)$ can only be computed in special cases.

Solution Group action is the correct generality for reparametrization
$\frac{\mathrm{d}}{\mathrm{d}g}\mathbb{E}_{q^g}[\ell] = \int_\Theta q_0(\theta)(\nabla_\theta \ell(g\cdot\theta))^\top \frac{\mathrm{d}g\cdot\theta}{\mathrm{d}g}\mathrm{d}\nu(\theta)$

[2] Mohamed et. al. *Monte carlo Gradient Estimation in Machine Learning* JMLR 2020

Solved issues

Issue 1	$\mathcal{Q}$ is required to be an exponential family.
Solution	Can choose q_0 freely and push it around with a group.
Issue 2	The updates could overshoot and leave the manifold.
Solution	Closure of the group under operation keeps updates on $\mathcal{Q}$.
Issue 3	The gradient $\nabla_\lambda \mathcal{E}(q_\lambda)$ can only be computed in special cases.
Solution	Group action is the correct generality for reparametrization $\frac{\mathrm{d}}{\mathrm{d}g}\mathbb{E}_{q^g}[\ell] = \int_\Theta q_0(\theta)(\nabla_\theta \ell(g \cdot \theta))^\top \frac{\mathrm{d}g \cdot \theta}{\mathrm{d}g} \mathrm{d}\nu(\theta)$
	Also called *pathwise gradient estimators*[2]

[2]Mohamed et. al. *Monte carlo Gradient Estimation in Machine Learning* JMLR 2020

Solved issues

Issue 1 $\mathcal{Q}$ is required to be an exponential family.

Solution Can choose q_0 freely and push it around with a group.

Issue 2 The updates could overshoot and leave the manifold.

Solution Closure of the group under operation keeps updates on $\mathcal{Q}$.

Issue 3 The gradient $\nabla_\lambda \mathcal{E}(q_\lambda)$ can only be computed in special cases.

Solution Group action is the correct generality for reparametrization
$\frac{\mathrm{d}}{\mathrm{d}g}\mathbb{E}_{q^g}[\ell] = \int_\Theta q_0(\theta)(\nabla_\theta \ell(g \cdot \theta))^\top \frac{\mathrm{d}g\cdot\theta}{\mathrm{d}g}\mathrm{d}\nu(\theta)$

Also called *pathwise gradient estimators*[2]

Bonus 1 The Fisher metric is invariant under translations by G.

[2]Mohamed et. al. *Monte carlo Gradient Estimation in Machine Learning* JMLR 2020

Solved issues

Issue 1 $\mathcal{Q}$ is required to be an exponential family.

Solution Can choose q_0 freely and push it around with a group.

Issue 2 The updates could overshoot and leave the manifold.

Solution Closure of the group under operation keeps updates on $\mathcal{Q}$.

Issue 3 The gradient $\nabla_\lambda \mathcal{E}(q_\lambda)$ can only be computed in special cases.

Solution Group action is the correct generality for reparametrization

$$\frac{\mathrm{d}}{\mathrm{d}g}\mathbb{E}_{q^g}[\ell] = \int_\Theta q_0(\theta)(\nabla_\theta \ell(g \cdot \theta))^\top \frac{\mathrm{d}g \cdot \theta}{\mathrm{d}g}\mathrm{d}\nu(\theta)$$

Also called *pathwise gradient estimators*[2]

Bonus 1 The Fisher metric is invariant under translations by G.

Bonus 2 The tangent directions Y at each step lie in the same vector space T_eG, so they can be accumulated from previous steps.

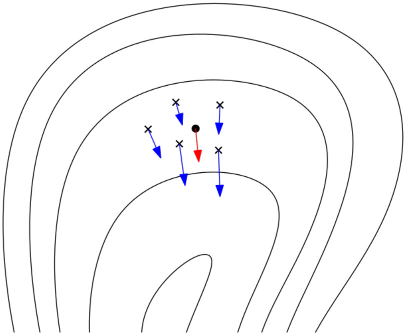
[2] Mohamed et. al. *Monte carlo Gradient Estimation in Machine Learning* JMLR 2020

Specific Update Formulas: The Additive Group

$$g \in \mathbb{R}^P \text{ additive} \qquad \Longrightarrow \qquad g \longleftarrow g - \alpha \mathbb{E}_{q_g}\left[\nabla_\theta \ell\right]$$

Specific Update Formulas: The Additive Group

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



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

$g \in \mathbb{R}_{>0}$ multiplicative $\implies$

$$g \longleftarrow g \exp\Big(- \alpha \big(\mathbb{E}_{q_g}[\theta \partial_\theta \ell] - \tau \big) \Big)$$

$(A, b) \in \mathrm{Aff}(\mathbb{R})$ affine group $\implies$

$$b \longleftarrow b + \frac{c_X}{c_y} A \frac{\exp(-\alpha U) - 1}{U} V$$

$$A \longleftarrow A \exp(-\alpha U)$$

$$\text{where} \qquad U = \mathbb{E}_{q_g}[(\theta - b)\partial_\theta \ell] - \tau$$

$$V = A\mathbb{E}_{q_g}[\partial_\theta \ell]$$

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 $j = 1, \ldots, K$

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

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

Filters of the multiplicative group

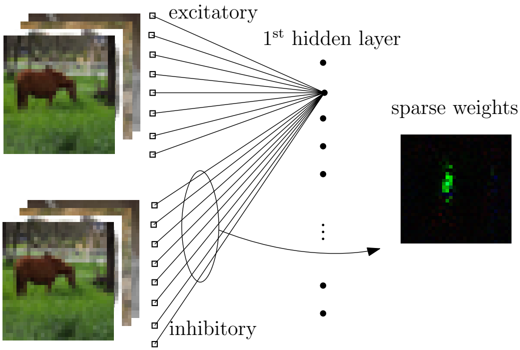
Label nodes in a neural network "excitatory" or "inhibitory" like biology.

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 $j = 1, \ldots, K$

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

$$g \leftarrow g \exp(-\alpha M)$$



Multiplicative vs Additive filters

Model & Dataset	**Method**	**Accuracy** ↑ (higher is better)	**NLL** ↓ (lower is better)	**ECE** ↓ (lower is better)
MNIST MLP	add.	$98.38_{\pm 0.02}$	$0.083_{\pm 0.001}$	$0.012_{\pm 0.000}$
	mult.	$98.59_{\pm 0.02}$	$0.058_{\pm 0.001}$	$0.006_{\pm 0.000}$
CIFAR–10 MLP	add.	$58.85_{\pm 0.08}$	$1.236_{\pm 0.002}$	$0.085_{\pm 0.001}$
	mult.	$59.19_{\pm 0.07}$	$1.160_{\pm 0.001}$	$0.026_{\pm 0.001}$

Additive rule is similar to SGD with momentum, multiplicative is different.

Multiplicative vs Additive filters

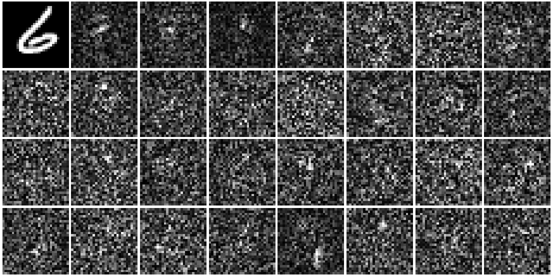
Model & Dataset	**Method**	**Accuracy** ↑ (higher is better)	**NLL** ↓ (lower is better)	**ECE** ↓ (lower is better)
MNIST MLP	add.	$98.38_{\pm 0.02}$	$0.083_{\pm 0.001}$	$0.012_{\pm 0.000}$
	mult.	$98.59_{\pm 0.02}$	$0.058_{\pm 0.001}$	$0.006_{\pm 0.000}$
CIFAR–10 MLP	add.	$58.85_{\pm 0.08}$	$1.236_{\pm 0.002}$	$0.085_{\pm 0.001}$
	mult.	$59.19_{\pm 0.07}$	$1.160_{\pm 0.001}$	$0.026_{\pm 0.001}$

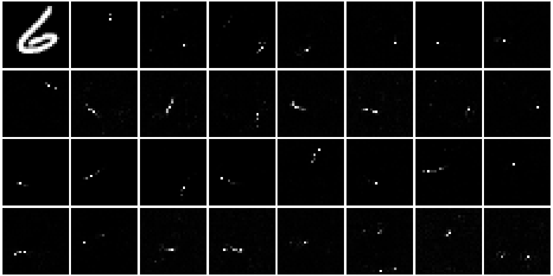
Additive rule is similar to SGD with momentum, multiplicative is different. They both learn.

Multiplicative vs Additive filters

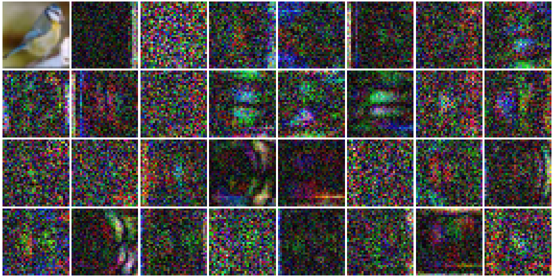
Model & Dataset	Method	Accuracy ↑ (higher is better)	NLL ↓ (lower is better)	ECE ↓ (lower is better)
MNIST MLP	add. mult.	$98.38_{\pm 0.02}$ $98.59_{\pm 0.02}$	$0.083_{\pm 0.001}$ $0.058_{\pm 0.001}$	$0.012_{\pm 0.000}$ $0.006_{\pm 0.000}$
CIFAR–10 MLP	add. mult.	$58.85_{\pm 0.08}$ $59.19_{\pm 0.07}$	$1.236_{\pm 0.002}$ $1.160_{\pm 0.001}$	$0.085_{\pm 0.001}$ $0.026_{\pm 0.001}$

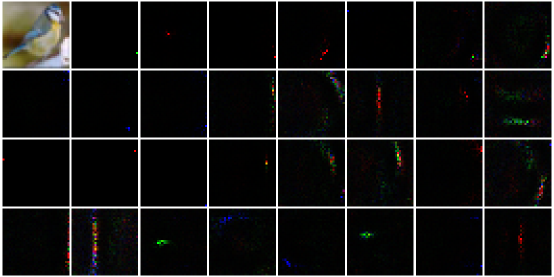
Additive rule is similar to SGD with momentum, multiplicative is different. They both learn.



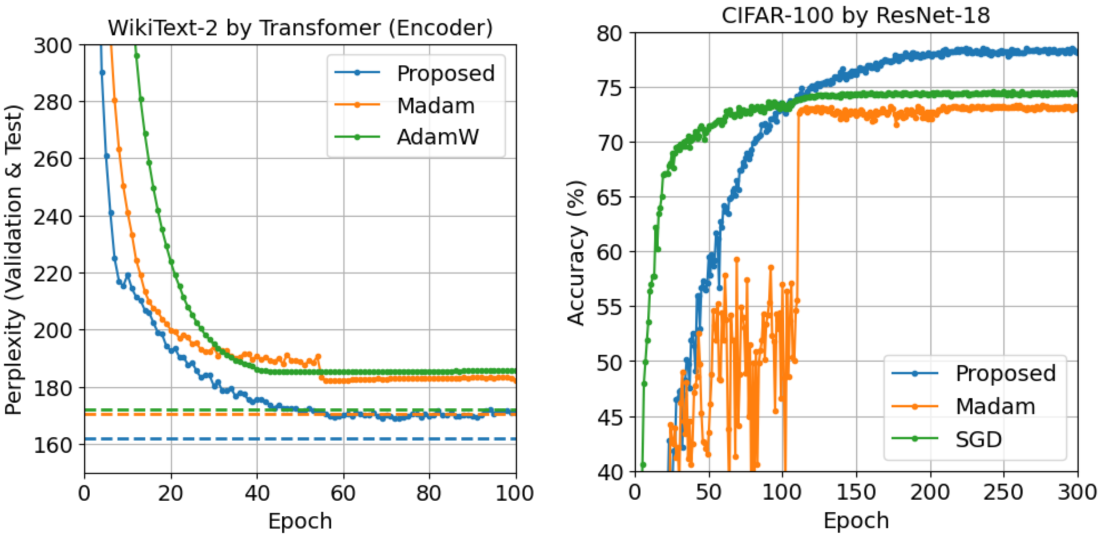


The additive vs multiplicative filters for RGB images





Further Results on the multiplicative update (modified) by Keigo Nishida



Exponential Families

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

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

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

Exponential Families

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

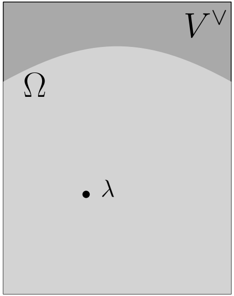
$$\Omega = \Omega_\nu(T) = \left\{ \lambda \in V^\vee : A(\lambda) := \log \int_\Theta e^{-\langle \lambda, T(\theta) \rangle} \mathrm{d}\nu(\theta) < \infty \right\}.$$

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

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

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

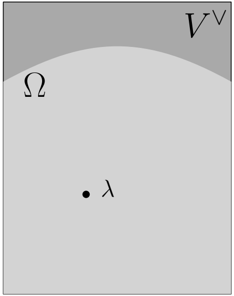
$$= \mathbb{E}_{q_\lambda}\left[\left(\tfrac{\partial}{\partial \lambda_i} \log q_\lambda\right)\left(\tfrac{\partial}{\partial \lambda_j} \log q_\lambda\right)\right] =: F_{i,j}(\lambda) \qquad \text{Fisher Matrix}$$



Exponential Families

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

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



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

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

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

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

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

Harmonic exponential families (by Koichi Tojo & Taro Yoshino)

What about exponential families on Θ 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 Θ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.
- ν 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 Θ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.
- ν 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 Θ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.
- ν 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 π 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 Θ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.
- ν 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 π 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 Θ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.
- ν 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 π 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 Θ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.
- ν 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 π 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 Θ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.
- ν 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 π 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 Θ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.
- ν 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 π 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 Θ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.
- ν 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 π 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 Θ which are closed under a Lie group action?

- Homogeneous space $\Theta \cong G/H$.
- ν 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 π 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 λ'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 λ's of exponential families. The overlap is harmonic exponential families:

Pushforwards of q_λ 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 λ's of exponential families. The overlap is harmonic exponential families:

Pushforwards of q_λ are still in the family,

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

Thus $\widetilde{\mathcal{Q}} = \{\lambda \in \Omega : \lambda = \pi^\vee(g)\lambda_0, g \in G\}$ 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

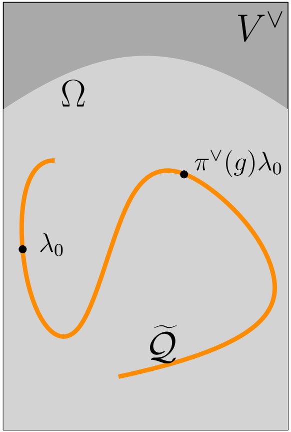
The Lie group rule is for transformation families. BLR is NGD on λ's of exponential families. The overlap is harmonic exponential families:

Pushforwards of q_λ 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_λ 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 λ's of exponential families. The overlap is harmonic exponential families:

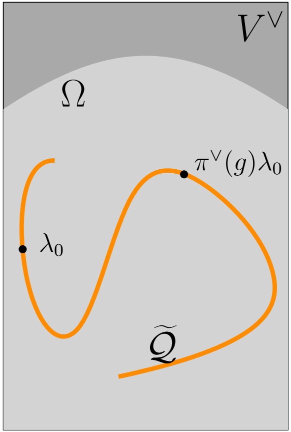
Pushforwards of q_λ are still in the family,

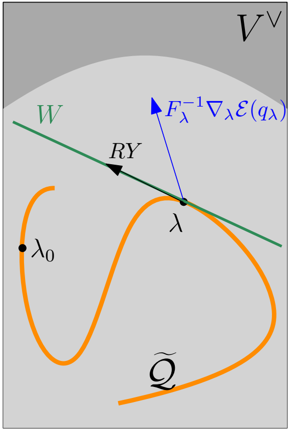
$$(q_\lambda)^g = q_{\lambda'} \quad \text{with} \quad \lambda' = \pi^\vee(g)\lambda$$

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_λ also vary with g:

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





Overview and Future work

- The Bayesian Learning Problem is a promising and fertile setting.

Overview and Future work

- The Bayesian Learning Problem is a promising and fertile setting.
- The BLR of Khan & Rue both specialized to many existing algorithms and gave rise to many new and successful algorithms.

Overview and Future work

- The Bayesian Learning Problem is a promising and fertile setting.
- The BLR of Khan & Rue both specialized to many existing algorithms and gave rise to many new and successful algorithms.
- Yet there are some issues with BLR such as: closure in the statistical manifold under updates, and calculation of derivatives w.r.t. distribution parameters.

Overview and Future work

- The Bayesian Learning Problem is a promising and fertile setting.
- The BLR of Khan & Rue both specialized to many existing algorithms and gave rise to many new and successful algorithms.
- Yet there are some issues with BLR such as: closure in the statistical manifold under updates, and calculation of derivatives w.r.t. distribution parameters.
- The group framework solves the closure problem by design and is able to very generally employ the **reparametrization trick**.

Overview and Future work

- The Bayesian Learning Problem is a promising and fertile setting.
- The BLR of Khan & Rue both specialized to many existing algorithms and gave rise to many new and successful algorithms.
- Yet there are some issues with BLR such as: closure in the statistical manifold under updates, and calculation of derivatives w.r.t. distribution parameters.
- The group framework solves the closure problem by design and is able to very generally employ the **reparametrization trick**.
- Each new group would deserve an empirical study to investigate their learning behaviours (like multiplicative vs. additive)

Overview and Future work

- The Bayesian Learning Problem is a promising and fertile setting.
- The BLR of Khan & Rue both specialized to many existing algorithms and gave rise to many new and successful algorithms.
- Yet there are some issues with BLR such as: closure in the statistical manifold under updates, and calculation of derivatives w.r.t. distribution parameters.
- The group framework solves the closure problem by design and is able to very generally employ the **reparametrization trick**.
- Each new group would deserve an empirical study to investigate their learning behaviours (like multiplicative vs. additive)
- There may be implementation problems with arbitrary Lie groups,
  e.g. the exponential map may not always be feasible to compute,
  so approximations may be necessary.

Teşekkürler

ありがとうございます

Vielen Danke

Merci

Thank you.

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 $\quad Y = \mathrm{Skew} \mathrm{Y}_0 = \frac{Y_0 - Y_0^\top}{2} \quad Y_0 = \mathbb{E}_{q_\Lambda}[\nabla \ell \theta^\top]$

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 $\quad Y = \mathrm{SkewY}_0 = \frac{Y_0 - Y_0^\top}{2} \quad Y_0 = \mathbb{E}_{q_\Lambda}[\nabla \ell \theta^\top]$

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

$$\Lambda \leftarrow e^{-\alpha Y}\Lambda$$

(actually an efficient variation is used)

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

G a Lie group $H \leq G$. Let ν be a relatively invariant measure on G $\pi : G \to \mathrm{GL}(V)$ a representation of G. Let α be a 1-cocycle of π 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 ν be a relatively invariant measure on Θ, meaning $\nu(gE) = \chi(g)\nu(E)$ for some homomorphism χ. 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 λ

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

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

Constrained maximization: Statistical mechanics interpretation

Assume θ to be a kind of "microstate" with energy level $\ell(\theta)$. So Θ 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 θ to be a kind of "microstate" with energy level $\ell(\theta)$. So Θ 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.

Why call it Bayesian?

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

Why call it Bayesian?

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

How to update p_τ? Take $\tau = 1$

Why call it Bayesian?

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

How to update p_τ? Take $\tau = 1$

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

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

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

Why call it Bayesian?

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

How to update p_τ? Take $\tau = 1$

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

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

Interpret $e^{-\ell_i(\theta)}$ as the likelihood of observing label y_i given the model parameter θ and $\mathbf{x}_i$. Interpret $\pi(\theta) \propto e^{-R(\theta)}$ 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 θ distribution. According to Bayes rule updated belief should be *after* a new data point.

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

Why call it Bayesian?

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

How to update p_τ? Take $\tau = 1$

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

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

Interpret $e^{-\ell_i(\theta)}$ as the likelihood of observing label y_i given the model parameter θ and $\mathbf{x}_i$.
Interpret $\pi(\theta) \propto e^{-R(\theta)}$ 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 θ distribution. According to Bayes rule updated belief should be *after* a new data point.

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

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