Meta-Strategy for Learning Tuning Parameters with Guarantees

Pierre Alquier – ABI team

RIKEN

AIP
Center for
Advanced Intelligence Project

AIP open seminar – 2021年3月10日

Joint work with Dimitri Meunier

Talk based on a joint work with :

Dimitri Meunier

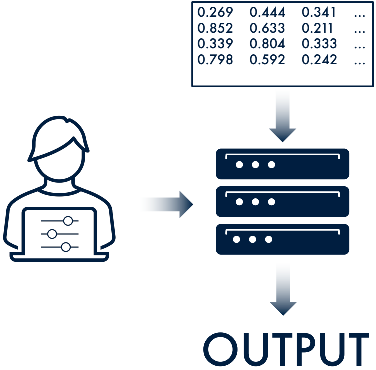
2020 : ENSAE Paris and ABI team

2021 : IIT Genoa

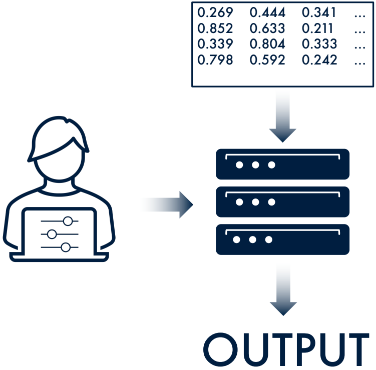
D. Meunier, P. Alquier (2021). Meta-Strategy for Learning Tuning Parameters with Guarantees. *Preprint arXiv :2102.02504*. Submitted.

Thank you to rc3 for the drawings in this talk.

Solving a task with an algorithm

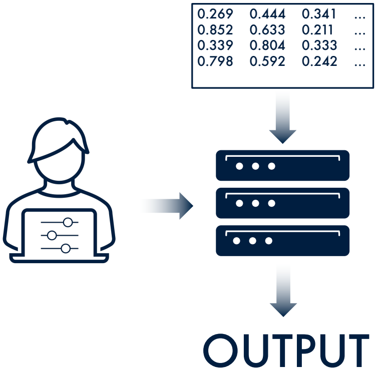


Solving a task with an algorithm



Examples :

Solving a task with an algorithm

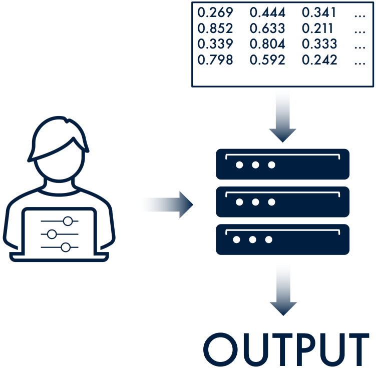


Examples :

- LASSO

$$\min_{\theta} \|y - X\theta\|^2 + \gamma \|\theta\|_1 .$$

Solving a task with an algorithm



Examples :

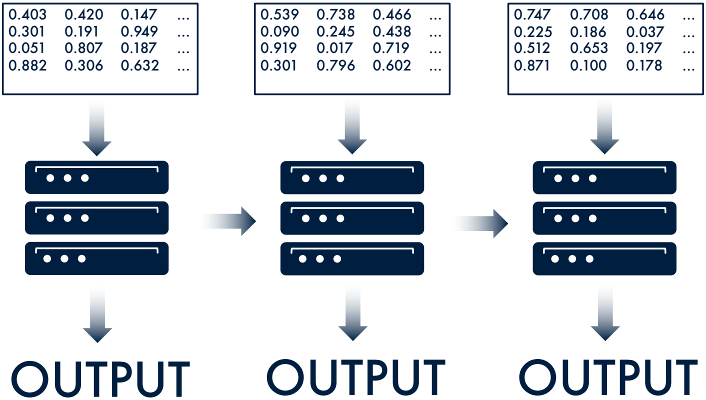
- LASSO

$$\min_{\theta} \|y - X\theta\|^2 + \gamma \|\theta\|_1.$$

- RIDGE

$$\min_{\theta} \|y - X\theta\|^2 + \alpha \|\theta\|_2^2.$$

Solving tasks sequentially to learn tuning parameters



A meta-strategy

Assume that we have an upper bound on the generalization error of a strategy when used with hyperparameter λ on task t :

$$\mathcal{L}(\text{data}_t, \lambda) = \mathcal{L}_t(\lambda).$$

A meta-strategy

Assume that we have an upper bound on the generalization error of a strategy when used with hyperparameter λ on task t :

$$\mathcal{L}(\mathrm{data}_t, \lambda) = \mathcal{L}_t(\lambda).$$

Idea : use online optimization algorithm to minimize the $\mathcal{L}_t$'s...

A meta-strategy

Assume that we have an upper bound on the generalization error of a strategy when used with hyperparameter λ on task t :

$$\mathcal{L}(\mathrm{data}_t, \lambda) = \mathcal{L}_t(\lambda).$$

Idea : use online optimization algorithm to minimize the $\mathcal{L}_t$'s...

Online Proximal Meta-Strategy (OPMS)

$$\lambda_{t+1} = \operatorname*{argmin}_{\lambda} \left\{ \mathcal{L}_t(\lambda) + \frac{\|\lambda - \lambda_t\|^2}{2\alpha} \right\}.$$

Example : tasks are online prediction

Sequential regression tasks $t = 1, 2, \ldots, T$

Example : tasks are online prediction

Sequential regression tasks $t = 1, 2, \ldots, T$

1. - $x_{t,1}$ given

Example : tasks are online prediction

Sequential regression tasks $t = 1, 2, \ldots, T$

1.
   - $x_{t,1}$ given
   - predict $y_{t,1}$: $f_{\theta_{t,1}}(x_{t,1})$

Example : tasks are online prediction

Sequential regression tasks $t = 1, 2, \ldots, T$

1.
   - $x_{t,1}$ given
   - predict $y_{t,1}$: $f_{\theta_{t,1}}(x_{t,1})$
   - $y_{t,1}$ is revealed

Example : tasks are online prediction

Sequential regression tasks $t = 1, 2, \ldots, T$

1.
   - $x_{t,1}$ given
   - predict $y_{t,1}$: $f_{\theta_{t,1}}(x_{t,1})$
   - $y_{t,1}$ is revealed
   - update $\theta_{t,1} \rightarrow \theta_{t,2}$

Example : tasks are online prediction

Sequential regression tasks $t = 1, 2, \ldots, T$

1.
   - $x_{t,1}$ given
   - predict $y_{t,1}$: $f_{\theta_{t,1}}(x_{t,1})$
   - $y_{t,1}$ is revealed
   - update $\theta_{t,1} \to \theta_{t,2}$
2.
   - $x_{t,2}$ given

Example : tasks are online prediction

Sequential regression tasks $t = 1, 2, \ldots, T$

1. - $x_{t,1}$ given
   - predict $y_{t,1}$: $f_{\theta_{t,1}}(x_{t,1})$
   - $y_{t,1}$ is revealed
   - update $\theta_{t,1} \to \theta_{t,2}$
2. - $x_{t,2}$ given
   - predict $y_{t,2}$: $f_{\theta_{t,2}}(x_{t,2})$

Example : tasks are online prediction

Sequential regression tasks $t = 1, 2, \ldots, T$

1.
   - $x_{t,1}$ given
   - predict $y_{t,1}$: $f_{\theta_{t,1}}(x_{t,1})$
   - $y_{t,1}$ is revealed
   - update $\theta_{t,1} \to \theta_{t,2}$
2.
   - $x_{t,2}$ given
   - predict $y_{t,2}$: $f_{\theta_{t,2}}(x_{t,2})$
   - $y_{t,2}$ revealed

Example : tasks are online prediction

Sequential regression tasks $t = 1, 2, \ldots, T$

1.
   - $x_{t,1}$ given
   - predict $y_{t,1}$: $f_{\theta_{t,1}}(x_{t,1})$
   - $y_{t,1}$ is revealed
   - update $\theta_{t,1} \to \theta_{t,2}$
2.
   - $x_{t,2}$ given
   - predict $y_{t,2}$: $f_{\theta_{t,2}}(x_{t,2})$
   - $y_{t,2}$ revealed
   - update $\theta_{t,2} \to \theta_{t,3}$

Example : tasks are online prediction

Sequential regression tasks $t = 1, 2, \ldots, T$

1.
   - $x_{t,1}$ given
   - predict $y_{t,1}$: $f_{\theta_{t,1}}(x_{t,1})$
   - $y_{t,1}$ is revealed
   - update $\theta_{t,1} \to \theta_{t,2}$
2.
   - $x_{t,2}$ given
   - predict $y_{t,2}$: $f_{\theta_{t,2}}(x_{t,2})$
   - $y_{t,2}$ revealed
   - update $\theta_{t,2} \to \theta_{t,3}$
3.
   - $x_{t,3}$ given

Example : tasks are online prediction

Sequential regression tasks $t = 1, 2, \ldots, T$

1.
   - $x_{t,1}$ given
   - predict $y_{t,1}$: $f_{\theta_{t,1}}(x_{t,1})$
   - $y_{t,1}$ is revealed
   - update $\theta_{t,1} \to \theta_{t,2}$
2.
   - $x_{t,2}$ given
   - predict $y_{t,2}$: $f_{\theta_{t,2}}(x_{t,2})$
   - $y_{t,2}$ revealed
   - update $\theta_{t,2} \to \theta_{t,3}$
3.
   - $x_{t,3}$ given
   - predict $y_{t,3}$: $f_{\theta_{t,3}}(x_{t,3})$

Example : tasks are online prediction

Sequential regression tasks $t = 1, 2, \dots, T$

1.
   - $x_{t,1}$ given
   - predict $y_{t,1}$: $f_{\theta_{t,1}}(x_{t,1})$
   - $y_{t,1}$ is revealed
   - update $\theta_{t,1} \to \theta_{t,2}$
2.
   - $x_{t,2}$ given
   - predict $y_{t,2}$: $f_{\theta_{t,2}}(x_{t,2})$
   - $y_{t,2}$ revealed
   - update $\theta_{t,2} \to \theta_{t,3}$
3.
   - $x_{t,3}$ given
   - predict $y_{t,3}$: $f_{\theta_{t,3}}(x_{t,3})$
   - $y_{t,3}$ revealed

Example : tasks are online prediction

Sequential regression tasks $t = 1, 2, \ldots, T$

1. - $x_{t,1}$ given
   - predict $y_{t,1}$: $f_{\theta_{t,1}}(x_{t,1})$
   - $y_{t,1}$ is revealed
   - update $\theta_{t,1} \to \theta_{t,2}$
2. - $x_{t,2}$ given
   - predict $y_{t,2}$: $f_{\theta_{t,2}}(x_{t,2})$
   - $y_{t,2}$ revealed
   - update $\theta_{t,2} \to \theta_{t,3}$
3. - $x_{t,3}$ given
   - predict $y_{t,3}$: $f_{\theta_{t,3}}(x_{t,3})$
   - $y_{t,3}$ revealed
   - update $\theta_{t,3} \to \theta_{t,4}$
4. ...

Example : tasks are online prediction

Sequential regression tasks $t = 1, 2, \dots, T$

1.
   - $x_{t,1}$ given
   - predict $y_{t,1}$: $f_{\theta_{t,1}}(x_{t,1})$
   - $y_{t,1}$ is revealed
   - update $\theta_{t,1} \to \theta_{t,2}$
2.
   - $x_{t,2}$ given
   - predict $y_{t,2}$: $f_{\theta_{t,2}}(x_{t,2})$
   - $y_{t,2}$ revealed
   - update $\theta_{t,2} \to \theta_{t,3}$
3.
   - $x_{t,3}$ given
   - predict $y_{t,3}$: $f_{\theta_{t,3}}(x_{t,3})$
   - $y_{t,3}$ revealed
   - update $\theta_{t,3} \to \theta_{t,4}$
4. ...

Loss function ℓ.

Example : tasks are online prediction

Sequential regression tasks $t = 1, 2, \ldots, T$

1. - $x_{t,1}$ given
   - predict $y_{t,1}$: $f_{\theta_{t,1}}(x_{t,1})$
   - $y_{t,1}$ is revealed
   - update $\theta_{t,1} \to \theta_{t,2}$
2. - $x_{t,2}$ given
   - predict $y_{t,2}$: $f_{\theta_{t,2}}(x_{t,2})$
   - $y_{t,2}$ revealed
   - update $\theta_{t,2} \to \theta_{t,3}$
3. - $x_{t,3}$ given
   - predict $y_{t,3}$: $f_{\theta_{t,3}}(x_{t,3})$
   - $y_{t,3}$ revealed
   - update $\theta_{t,3} \to \theta_{t,4}$
4. ...

Loss function ℓ.
For short,

$$\ell_{t,i}(\theta) = \ell(y_{t,i}, f_\theta(x_{t,i})).$$

Example : tasks are online prediction

Sequential regression tasks $t = 1, 2, \ldots, T$

1. - $x_{t,1}$ given
   - predict $y_{t,1} : f_{\theta_{t,1}}(x_{t,1})$
   - $y_{t,1}$ is revealed
   - update $\theta_{t,1} \to \theta_{t,2}$
2. - $x_{t,2}$ given
   - predict $y_{t,2} : f_{\theta_{t,2}}(x_{t,2})$
   - $y_{t,2}$ revealed
   - update $\theta_{t,2} \to \theta_{t,3}$
3. - $x_{t,3}$ given
   - predict $y_{t,3} : f_{\theta_{t,3}}(x_{t,3})$
   - $y_{t,3}$ revealed
   - update $\theta_{t,3} \to \theta_{t,4}$
4. ...

Loss function ℓ.
For short,

$$\ell_{t,i}(\theta) = \ell(y_{t,i}, f_\theta(x_{t,i})).$$

Objective :

$$\sum_{i=1}^{n} \ell_{t,i}(\theta_{t,i})$$

as small as possible.

Example : online gradient algorithm (OGA)

$$\theta_{t,i+1} = \theta_{t,i} - \eta \nabla \ell_{t,i}(\theta_{t,i})$$

Example : online gradient algorithm (OGA)

$$\theta_{t,i+1} = \theta_{t,i} - \eta \nabla \ell_{t,i}(\theta_{t,i})$$

Regret bound for OGA

If each $\ell_{t,i}$ is convex and L-Lipschitz,

$$\sum_{i=1}^{n} \ell_{t,i}(\theta_{t,i}) \leq \underbrace{\inf_{\|\theta\| \leq B} \left\{ \sum_{i=1}^{n} \ell_{t,i}(\theta) + \frac{\eta n L^2}{2} + \frac{\|\theta - \theta_{1,t}\|^2}{2\eta} \right\}}_{\mathcal{L}_t(\eta,\theta_{t,1}) = \mathcal{L}_t(\lambda)}.$$

Example : online gradient algorithm (OGA)

$$\theta_{t,i+1} = \theta_{t,i} - \eta \nabla \ell_{t,i}(\theta_{t,i})$$

Regret bound for OGA

If each $\ell_{t,i}$ is convex and L-Lipschitz,

$$\sum_{i=1}^{n} \ell_{t,i}(\theta_{t,i}) \leq \underbrace{\inf_{\|\theta\| \leq B} \left\{ \sum_{i=1}^{n} \ell_{t,i}(\theta) + \frac{\eta n L^2}{2} + \frac{\|\theta - \theta_{1,t}\|^2}{2\eta} \right\}}_{\mathcal{L}_t(\eta,\theta_{t,1}) = \mathcal{L}_t(\lambda)}.$$

Take $\eta \sim 1/\sqrt{n}$ to get :

$$\sum_{i=1}^{n} \ell_{t,i}(\theta_{t,i}) \leq \inf_{\|\theta\| \leq B} \sum_{i=1}^{n} \ell_{t,i}(\theta) + \mathcal{O}(\sqrt{n}).$$

Example : online gradient algorithm (OGA)

$$\theta_{t,i+1} = \theta_{t,i} - \eta \nabla \ell_{t,i}(\theta_{t,i})$$

Regret bound for OGA

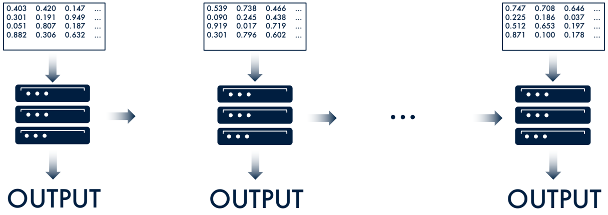
If each $\ell_{t,i}$ is convex and L-Lipschitz,

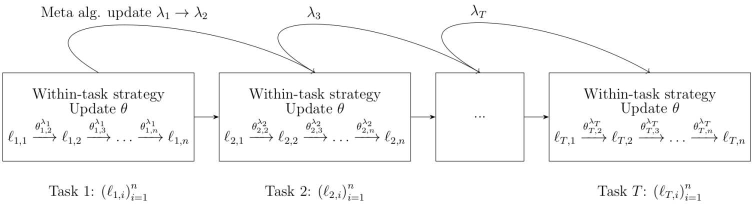
$$\sum_{i=1}^{n} \ell_{t,i}(\theta_{t,i}) \leq \underbrace{\inf_{\|\theta\|\leq B} \left\{ \sum_{i=1}^{n} \ell_{t,i}(\theta) + \frac{\eta n L^2}{2} + \frac{\|\theta - \theta_{1,t}\|^2}{2\eta} \right\}}_{\mathcal{L}_t(\eta,\theta_{t,1}) = \mathcal{L}_t(\lambda)}.$$

Take $\eta \sim 1/\sqrt{n}$ to get :

$$\frac{1}{n} \sum_{i=1}^{n} \ell_{t,i}(\theta_{t,i}) \leq \inf_{\|\theta\|\leq B} \frac{1}{n} \sum_{i=1}^{n} \ell_{t,i}(\theta) + \mathcal{O}\left(\frac{1}{\sqrt{n}} \right).$$

Meta-learning for OGA





Meta-learning for OGA

Online Proximal Meta-Strategy (OPMS)

$$\lambda_{t+1} = \underset{\lambda}{\text{argmin}} \left\{ \mathcal{L}_t(\lambda) + \frac{\|\lambda - \lambda_t\|^2}{2\alpha} \right\}.$$

Meta-learning for OGA

Online Proximal Meta-Strategy (OPMS)

$$\lambda_{t+1} = \operatorname*{argmin}_{\lambda} \left\{ \mathcal{L}_t(\lambda) + \frac{\|\lambda - \lambda_t\|^2}{2\alpha} \right\}.$$

$$\begin{aligned}
(\eta_{t+1}, \theta_{t+1,1}) \\
&= \operatorname*{argmin}_{\eta,\vartheta} \min_{\|\theta\| \leq B} \left\{ \sum_{i=1}^{n} \ell_{t,i}(\theta) + \frac{\eta n L^2}{2} + \frac{\|\theta - \vartheta\|^2}{2\eta} \right. \\
&\qquad \left. + \frac{\|\vartheta - \theta_{t,1}\|^2 + (\eta - \eta_t)^2}{2\alpha} \right\}.
\end{aligned}$$

The big question : what do we win ?

Question : what do we win ?

Standard : learning in isolation

$$\frac{1}{n}\sum_{i=1}^{n}\ell_{1,i}(\theta_{1,i}) \leq \inf_{\|\theta\|\leq B}\frac{1}{n}\sum_{i=1}^{n}\ell_{1,i}(\theta) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$$

$$+$$

$$\vdots$$

$$+$$

$$\frac{1}{n}\sum_{i=1}^{n}\ell_{T,i}(\theta_{T,i}) \leq \inf_{\|\theta\|\leq B}\frac{1}{n}\sum_{i=1}^{n}\ell_{T,i}(\theta) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$$

Standard : learning in isolation

$$\frac{1}{n}\sum_{i=1}^{n}\ell_{1,i}(\theta_{1,i}) \leq \inf_{\|\theta\|\leq B} \frac{1}{n}\sum_{i=1}^{n}\ell_{1,i}(\theta) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$$

$$+$$

$$\vdots$$

$$+$$

$$\frac{1}{n}\sum_{i=1}^{n}\ell_{T,i}(\theta_{T,i}) \leq \inf_{\|\theta\|\leq B} \frac{1}{n}\sum_{i=1}^{n}\ell_{T,i}(\theta) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$$

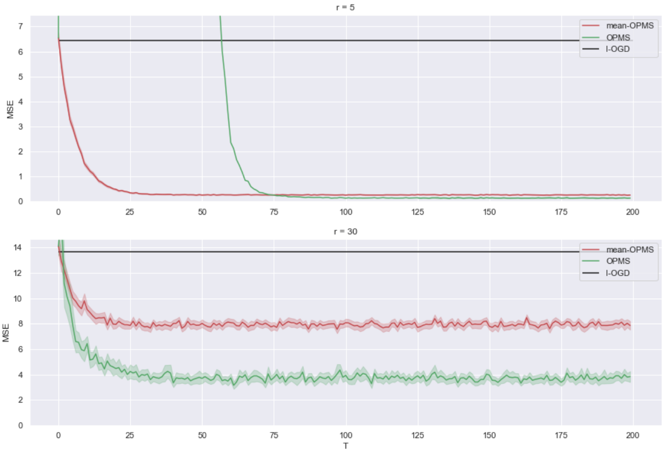
$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\ell_{t,i}(\theta_{1,i}) \leq \inf_{\substack{\|\theta_1\|\leq B\\ \dots \\ \|\theta_T\|\leq B}} \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\ell_{t,i}(\theta_t) + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

Guarantees for our meta-strategy

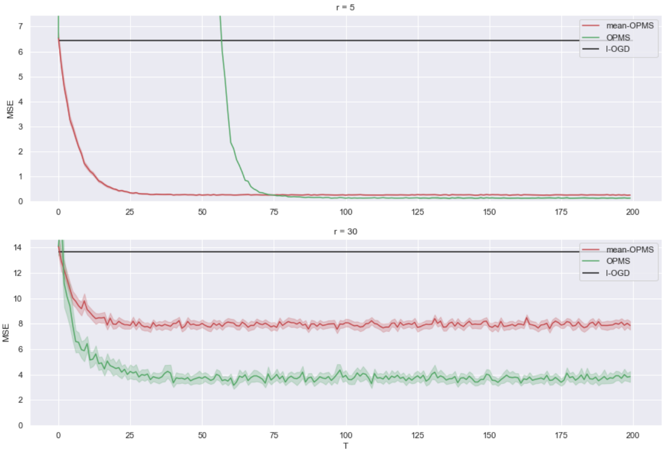
Theorem

$$\frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\ell_{t,i}(\theta_{1,i}) \leq \inf_{\substack{\|\theta_1\| \leq B \\ \dots \\ \|\theta_T\| \leq B}} \left\{ \frac{1}{nT}\sum_{t=1}^{T}\sum_{i=1}^{n}\ell_{t,i}(\theta_t) + \mathcal{O}\left(\frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\theta_t - \bar{\theta}\right)^2}}{\sqrt{n}} + \frac{1}{n} + \frac{n}{\sqrt{T}} \right) \right\}.$$

Simulated examples



Simulated examples



r=5
6.44
0.27
0.15
r=30
13.60
7.93
3.72

But....

Laboratories

TOP ▸ Laboratories ▸ Generic Technology Research Group ▸ Approximate Bayesian Inference Team

Laboratory

いいね！ Tweet

Approximate Bayesian Inference Team

Mohammad Emtiyaz Khan
(Ph.D.)

Title

Team Leader

Members

Team leader Mohammad Emtiyaz Khan	Research Scientist Pierre Alain Alquier
Postdoctoral researcher Thomas Moellenhoff	Postdoctoral researcher Gian Maria Marconi
Technical Staff I Dharmesh Vijay Tailor	

But.... but .....

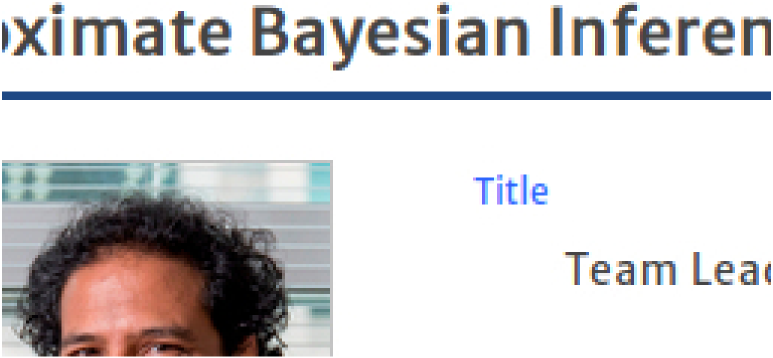
Approximate Bayesian Inference Team

Mohammad Emtiyaz Khan
(Ph.D.)

Title

Team Leader

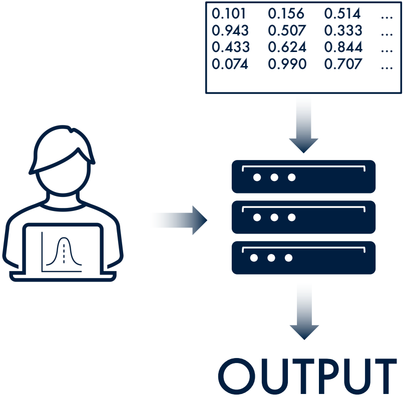
But.... but ..... but....



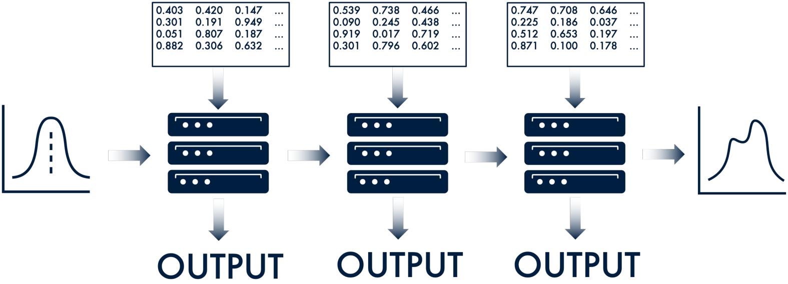
But !

Where is Bayes ? ?

Bayesian inference



Learning the prior



Online variational inference

B.-E. Chérief-Abdellatif, P. Alquier, M. E. Khan (2019). *A generalization bound for online variational inference.* ACML.

Online variational inference :

$$\mu_{t,i} = \operatorname*{argmin}_{\mu \in M} \left\{ \mu^T \sum_{j=1}^{i-1} \nabla_{\mu_{t,j}} \mathbb{E}_{\theta \sim q_{\mu_{t,j}}} [\ell_{t,j}(\theta)] + \frac{\mathcal{K}(q_\mu, \pi)}{\eta} \right\}.$$

Online variational inference

B.-E. Chérief-Abdellatif, P. Alquier, M. E. Khan (2019). *A generalization bound for online variational inference.* ACML.

Online variational inference :

$$\mu_{t,i} = \operatorname*{argmin}_{\mu \in M} \left\{ \mu^T \sum_{j=1}^{i-1} \nabla_{\mu_{t,j}} \mathbb{E}_{\theta \sim q_{\mu_{t,j}}} [\ell_{t,j}(\theta)] + \frac{\mathcal{K}(q_\mu, \pi)}{\eta} \right\}.$$

Regret bound :

$$\sum_{i=1}^{n} \mathbb{E}_{\theta \sim q_{\mu_{t,i}}} [\ell_{t,i}(\theta)] \leq \inf_{\mu \in \mathcal{M}} \left\{ \mathbb{E}_{\theta \sim q_\mu} \left[\sum_{i=1}^{n} \ell_{t,i}(\theta) \right] + \frac{\eta 4 L^2 n}{\alpha} + \frac{\mathcal{K}(q_\mu, \pi)}{\eta} \right\}.$$

Online variational inference

B.-E. Chérief-Abdellatif, P. Alquier, M. E. Khan (2019). *A generalization bound for online variational inference.* ACML.

Online variational inference :

$$\mu_{t,i} = \operatorname*{argmin}_{\mu \in M} \left\{ \mu^T \sum_{j=1}^{i-1} \nabla_{\mu_{t,j}} \mathbb{E}_{\theta \sim q_{\mu_{t,j}}} [\ell_{t,j}(\theta)] + \frac{\mathcal{K}(q_\mu, \pi)}{\eta} \right\}.$$

Regret bound :

$$\sum_{i=1}^{n} \mathbb{E}_{\theta \sim q_{\mu_{t,i}}} [\ell_{t,i}(\theta)] \leq \inf_{\mu \in \mathcal{M}} \left\{ \mathbb{E}_{\theta \sim q_\mu} \left[\sum_{i=1}^{n} \ell_{t,i}(\theta) \right] + \frac{\eta 4 L^2 n}{\alpha} + \frac{\mathcal{K}(q_\mu, \pi)}{\eta} \right\}.$$

If $q_\mu = \mathcal{N}(\mu, I)$ and $\pi = \mathcal{N}(m, I)$, $\mathcal{K}(q_\mu, \pi) = \frac{\|\mu - m\|^2}{2}$.

どうも ありがとう ございました!