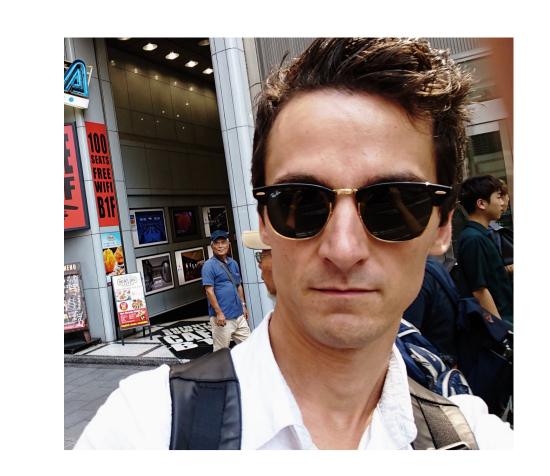
Non-Exponentially Weighted Aggregation: Regret Bounds for Unbounded Loss Functions

Pierre Alquier RIKEN AIP





(Generalized) Bayes update

$$oldsymbol{
ho}^t = \mathop{\mathsf{arg\,min}}_{oldsymbol{
ho}} \left\{ \sum_{s=1}^{t-1} \mathbb{E}_{ heta \sim
ho}[\ell_s(oldsymbol{ heta})] + rac{\mathrm{KL}(oldsymbol{
ho} \| oldsymbol{\pi})}{oldsymbol{\eta}}
ight\}.$$

 \blacktriangleright no constraint on ρ :

$$ho^t(\mathrm{d} heta) \propto \exp\left[-\eta \sum_{s=1}^{t-1} \ell_s(heta)
ight] \pi(\mathrm{d} heta).$$

 \triangleright constraint on ρ : variational inference.

Reasons to go beyond \mathbf{KL} :

KNOBLAUCH, J., JEWSON, J. & DAMOULAS, T. (2019). Generalized variational inference: Three arguments for deriving new posteriors. *Preprint arXiv*.

Objective

$$ho^t = rg \min_{
ho} \left\{ \sum_{s=1}^{t-1} \mathbb{E}_{ heta \sim
ho}[\ell_s(heta)] + rac{\mathcal{D}_{\phi}(
ho \| \pi)}{\eta}
ight\}.$$
 $\mathcal{D}_{\phi}(
ho \| \pi) = \mathbb{E}_{ heta \sim \pi} \left[\phi \left(rac{\mathrm{d}
ho}{\mathrm{d} \pi}(heta)
ight)
ight]$

- formula for the update?
- regret bounds?

Theorem: formula for ρ^t – "non-exponential weights"

$$egin{aligned} oldsymbol{
abla} ilde{\phi}^*(y) &= rg \max_{x \geq 0} \left\{ xy - \phi(x)
ight\}, \ & \sum_{s \geq 0} \ell_s(heta) = oldsymbol{
abla} ilde{\phi}^* \left(oldsymbol{\lambda}_t - \eta \sum_{s = 1}^{t-1} \ell_s(heta)
ight) \pi(heta heta). \end{aligned}$$

The proof uses convex analysis tools from:

AGRAWAL, R. & HOREL, T. (2020). Optimal Bounds between *f*-Divergences and Integral Probability Metrics. *ICML*.

Example 1: $D_\phi(ho\|\pi)=\mathrm{KL}(ho\|\pi)$

$$egin{aligned} \Phi(x) &= x \log(x) \
abla ilde{\phi}^*(y) &= \exp(y) - 1 \end{aligned}$$
 $ho^t(\mathrm{d} heta) &= \exp\left[oldsymbol{\lambda}_t - \eta \sum_{s=1}^{t-1} \ell_s(heta) - 1
ight] \pi(\mathrm{d} heta)$

Example 2: $D_{\phi}(ho\|\pi)=\chi^2(ho\|\pi)$

$$\Phi(x) = x^2 - 1$$
 $\nabla \tilde{\phi}^*(y) = \max(0, y/2)$
 $ho^t(\mathrm{d}\theta) = rac{1}{2} \max \left[0, \lambda_t - \eta \sum_{s=1}^{t-1} \ell_s(\theta)\right] \pi(\mathrm{d}\theta)$

Theorem: regret bound

Assume there is a norm $\|\cdot\|$ such that

- 1. $\rho \mapsto \mathbb{E}_{\theta \sim \rho}[\ell_t(\theta)]$ is *L*-Lipschitz w.r.t $\|\cdot\|$,
- 2. $\rho \mapsto D_{\phi}(\rho \| \pi)$ is α -strongly convex w.r.t $\| \cdot \|$.

$$\sum_{t=1}^T \mathbb{E}_{\theta \sim \rho^t}[\ell_t(\theta)] \leq \inf_{\rho \in \mathcal{P}(\Theta)} \left\{ \sum_{t=1}^T \mathbb{E}_{\theta \sim \rho}[\ell_t(\theta)] + \frac{\eta L^2 T}{\alpha} + \frac{D_{\phi}(\rho \| \pi)}{\eta} \right\}.$$

Example 1: $D_{\phi}(ho\|\pi) = \mathrm{KL}(ho\|\pi)$

- ▶ known result: $KL(\rho || \pi)$ is 1-strongly convex with respect to $|| \cdot ||_{TV}$;
- we have:

$$\left| \int \ell_t(\theta) \rho(\mathrm{d}\theta) - \int \ell_t \rho'(\mathrm{d}\theta) \right| \leq \int \ell_t(\theta) \left| \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) - \frac{\mathrm{d}\rho'}{\mathrm{d}\pi}(\theta) \right| \pi(\mathrm{d}\theta)$$

$$\leq L \int \left| \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) - \frac{\mathrm{d}\rho'}{\mathrm{d}\pi}(\theta) \right| \pi(\mathrm{d}\theta)$$

$$= 2\|\rho - \rho'\|_{\mathrm{TV}}$$

on the condition that $0 \leq \ell_t(\theta) \leq L$ for any θ .

Example 2: $D_{\phi}(ho\|\pi)=\chi^2(ho\|\pi)$

- $\blacktriangleright \phi(x) = x^2 1$ is 2-strongly convex so D_{ϕ} is 2-strongly convex with respect to the $L_2(\pi)$ norm.
- ▶ we have

$$\left| \int \ell_t(\theta) \rho(\mathrm{d}\theta) - \int \ell_t \rho'(\mathrm{d}\theta) \right| \leq \int \ell_t(\theta) \left| \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) - \frac{\mathrm{d}\rho'}{\mathrm{d}\pi}(\theta) \right| \pi(\mathrm{d}\theta)$$

$$\leq L \left(\int \left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) - \frac{\mathrm{d}\rho'}{\mathrm{d}\pi}(\theta) \right)^2 \pi(\mathrm{d}\theta) \right)^{1/2}$$

on the condition that $\left(\int \ell_t(\theta)^2 \pi(\mathrm{d}\theta)\right)^{1/2} \leq L$.

Constrained optimization

Constraint: $ho\in\mathcal{F}=\{q_\mu,\mu\in M\}$ a parametric family. Example: Gaussian distributions.

Initial objective:

$$m{\mu}^t = rg\min_{m{\mu} \in \mathcal{M}} \left\{ \sum_{s=1}^{t-1} \mathbb{E}_{m{ heta} \sim m{q}_{m{\mu}}} [\ell_s(m{ heta})] + rac{D_{m{\phi}}(m{q}_{m{\mu}},m{\pi})}{m{\eta}}
ight\}.$$

Linearization gives:

$$oldsymbol{\mu}^t = rg\min_{oldsymbol{\mu} \in \mathcal{M}} \left\{ \sum_{s=1}^{t-1} ig\langle oldsymbol{\mu}, oldsymbol{
abla} \mathbb{E}_{oldsymbol{ heta} \sim q_{oldsymbol{\mu}^s}} [\ell_s(oldsymbol{ heta})] ig
angle + rac{D_\phi(q_\mu, oldsymbol{\pi})}{oldsymbol{\eta}}
ight\}$$

Explicit update

$$egin{aligned} m{F}(\mu) &:= D_\phi(q_\mu, \pi) \ \ \mu_t &= m{
abla} m{F}^* \left(-\eta \sum_{s=1}^{t-1} m{
abla}_{\mu=\mu_s} \mathbb{E}_{ heta \sim q_\mu} [\ell_s(heta)]
ight). \end{aligned}$$

Mirror descent structure: initialize $\lambda_0=0$, and update at each step:

$$egin{cases} oldsymbol{\lambda}_t = oldsymbol{\lambda}_{t-1} - \eta
abla_{\mu=\mu_{t-1}} \mathbb{E}_{ heta \sim q_{\mu}} [\ell_{t-1}(heta)], \ oldsymbol{\mu}_t =
abla F^*(oldsymbol{\lambda}_t) \end{cases}$$

Regret bound

Let $\|\cdot\|$ be a norm on \mathbb{R}^d . If each $\mu \mapsto \mathbb{E}_{\theta \sim q_{\mu}}[\ell_s(\theta)]$ is convex and L-Lipschitz with respect to $\|\cdot\|$, if $\mu \mapsto D_{\phi}(q_{\mu}||\pi)$ is α -strongly convex with respect to $\|\cdot\|$,

$$\left|\sum_{t=1}^T \mathbb{E}_{\theta \sim q_{\mu_t}}[\ell_t(\theta)] \leq \inf_{\mu \in \mathcal{M}} \left\{ \sum_{t=1}^T \mathbb{E}_{\theta \sim q_{\mu}}[\ell_t(\theta)] + \frac{\eta L^2 T}{\alpha} + \frac{D_{\phi}(q_{\mu}||\pi)}{\eta} \right\}$$

Example where D_ϕ is strongly convex: Gaussian family, \mathbf{KL} case, studied in

© CHÉRIEF-ABDELLATIF, B.-E., ALQUIER, P. & KHAN, M. E. (2019). A generalization bound for online variational inference. *ACML*.

Conditions on the expected loss studied in

DOMKE, J. (2020). Provable smoothness guarantees for black-box variational inference. *ICML*.

The proof is based on an adaptation of the study of FTRL, see e.g.:

SHALEV-SHWARTZ, S. (2011). Online learning and online convex optimization. Foundations and trends in Machine Learning.