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

Pierre Alquier

RIKEN AIP

RIKEN

AIP Center for Advanced Intelligence Project

(Generalized) Bayes update

$$\rho^t = \arg\min_{\rho} \left\{ \sum_{s=1}^{t-1} \mathbb{E}_{\theta\sim\rho}[\ell_s(\theta)] + \frac{\mathbf{KL}(\rho\|\pi)}{\eta} \right\}.$$

- no constraint on ρ:

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

- 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

$$\rho^t = \arg\min_{\rho} \left\{ \sum_{s=1}^{t-1} \mathbb{E}_{\theta\sim\rho}[\ell_s(\theta)] + \frac{D_\phi(\rho\|\pi)}{\eta} \right\}.$$

$$D_\phi(\rho\|\pi) = \mathbb{E}_{\theta\sim\pi}\left[\phi\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) \right) \right]$$

- formula for the update?
- regret bounds?

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

$$\nabla\tilde{\phi}^*(y) = \arg\max_{x\geq 0} \{xy - \phi(x)\},$$

$$\rho^t(\mathrm{d}\theta) = \nabla\tilde{\phi}^*\left(\lambda_t - \eta \sum_{s=1}^{t-1} \ell_s(\theta) \right) \pi(\mathrm{d}\theta).$$

The proof uses convex analysis tools from:

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

Example 1: $D_\phi(\rho\|\pi) = \mathbf{KL}(\rho\|\pi)$

$$\Phi(x) = x\log(x)$$
$$\nabla\tilde{\phi}^*(y) = \exp(y) - 1$$

$$\rho^t(\mathrm{d}\theta) = \exp\left[\lambda_t - \eta \sum_{s=1}^{t-1} \ell_s(\theta) - 1 \right] \pi(\mathrm{d}\theta)$$

Example 2: $D_\phi(\rho\|\pi) = \chi^2(\rho\|\pi)$

$$\Phi(x) = x^2 - 1$$
$$\nabla\tilde{\phi}^*(y) = \max(0, y/2)$$

$$\rho^t(\mathrm{d}\theta) = \frac{1}{2}\max\left[0, \lambda_t - \eta \sum_{s=1}^{t-1} \ell_s(\theta) \right] \pi(\mathrm{d}\theta)$$

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: $D_\phi(\rho\|\pi) = \mathbf{KL}(\rho\|\pi)$

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

$$\left| \int \ell_t(\theta)\rho(\mathrm{d}\theta) - \int \ell_t \rho'(\mathrm{d}\theta) \right| \leq \int \ell_t(\theta) \left| \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) - \frac{\mathrm{d}\rho'}{\mathrm{d}\pi}(\theta) \right| \pi(\mathrm{d}\theta)$$
$$\leq L \underbrace{\int \left| \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\theta) - \frac{\mathrm{d}\rho'}{\mathrm{d}\pi}(\theta) \right| \pi(\mathrm{d}\theta)}_{=2\|\rho-\rho'\|_{\mathrm{TV}}}$$

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

Example 2: $D_\phi(\rho\|\pi) = \chi^2(\rho\|\pi)$

- $\phi(x) = x^2 - 1$ 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: $\rho \in \mathcal{F} = \{q_\mu, \mu \in M\}$ a parametric family. Example: Gaussian distributions.

Initial objective:

$$\mu^t = \arg\min_{\mu\in M} \left\{ \sum_{s=1}^{t-1} \mathbb{E}_{\theta\sim q_\mu}[\ell_s(\theta)] + \frac{D_\phi(q_\mu,\pi)}{\eta} \right\}.$$

Linearization gives:

$$\mu^t = \arg\min_{\mu\in M} \left\{ \sum_{s=1}^{t-1} \langle \mu, \nabla\mathbb{E}_{\theta\sim q_{\mu^s}}[\ell_s(\theta)] \rangle + \frac{D_\phi(q_\mu,\pi)}{\eta} \right\}.$$

Explicit update

$$F(\mu) := D_\phi(q_\mu,\pi)$$

$$\mu_t = \nabla F^*\left(-\eta \sum_{s=1}^{t-1} \nabla_{\mu=\mu_s}\mathbb{E}_{\theta\sim q_\mu}[\ell_s(\theta)] \right).$$

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

$$\begin{cases} \lambda_t = \lambda_{t-1} - \eta\nabla_{\mu=\mu_{t-1}}\mathbb{E}_{\theta\sim q_\mu}[\ell_{t-1}(\theta)], \\ \mu_t = \nabla F^*(\lambda_t) \end{cases}$$

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 $\|\cdot\|$,

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

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.*