Scale-free Unconstrained Online Learning for Curved Losses

†

Jack Mayo, Hédi Hadiji†, Tim van Erven

Dutch Seminar on Optimzation, December 8th, 2022

Setting: Online Supervised Learning

For $t = 1, 2, \ldots, T$

- Receive feature $x_t \in \mathcal{X}$
- Play action $a_t \in \mathcal{A}$
- Receive loss $\ell(a_t, y_t)$ with $y_t \in \mathcal{Y}$

Performance against $\mathcal{F} = \{f_\theta : \mathcal{X} \to \mathcal{A} \mid \theta \in \Theta\}$ measured by

$$R_T(\theta) = \sum_{t=1}^{T} \ell(a_t, y_t) - \sum_{t=1}^{T} \ell(f_\theta(x_t), y_t) \quad \text{for } \theta \in \Theta$$

Online Convex Optimization:

- Assume $\theta \mapsto \ell_t(\theta) := \ell(f_\theta(x_t), y_t)$ convex and $\Theta \subseteq \mathbb{R}^d$
- Play *parameter* $\theta_t \in \Theta$

Adaptivity to Gradients and Comparator in OCO

Two main goals:

- Adapt to $\|\theta\|$ (comparator norm)
- Adapt to $G = \max_{t \in [T]} \|\nabla \ell_t(\theta_t)\|$ (gradient length/data range)

- $U \geqslant \|\theta\|$ known, G (possibly) unknown: [Zinkevich '03, Duchi et al. '11]

$$R_T(\theta) = \mathcal{O}\big(UG\sqrt{T}\big)$$

- G known, U unknown: [McMahan and Streeter '12]

$$R_T(\theta) = \mathcal{O}\big(\|\theta\| G \sqrt{T \log(1 + \|\theta\| T)}\,\big)$$

- Both G and U unknown: [Cutkosky '19, Mhammedi and Koolen '20]

$$R_T(\theta) = \mathcal{O}\big(\|\theta\| G \sqrt{T \log(1 + \|\theta\| T)} + G\|\theta\|^3\big)$$

Price for adaptivity!

Plot Twist: Adaptivity for Free in Online Supervised Learning

1-Lipschitz losses, linear model $f_\theta(x) = \theta^\intercal x$ (e.g. Hinge loss)
[Kempka et al. '19, Mhammedi, Koolen '20]:

- $\|\nabla \ell_t(\theta_t)\| \leqslant \|x_t\|$
- Adapt to both $\|\theta\|$ and $X = \max \|x_t\|$ **almost for free**

$$R_T(\theta) = \mathcal{O}\big(\|\theta\| X \sqrt{T \log(\|\theta\| X T)}\,\big)$$

- Scale-free algorithms get the right dependence on X

Q: *For other losses, what is the cost of adapting to* $\|\theta\|$ *and the data range?*

A: In many cases, free!

Approach

- **Key property:** η-*Mixability* of the loss ℓ:

Definition (η-Mixability)

A loss $\ell : \mathcal{A} \to \mathbb{R}$ is called η-*mixable* if, for some $\eta > 0$ and all $p \in \mathcal{P}_{\mathcal{A}}$ there exists some $\zeta : \mathcal{P}_{\mathcal{A}} \to \mathcal{A}$ such that

$$\ell(\zeta(p)) \leqslant -\frac{1}{\eta} \ln \mathbb{E}_{a \sim p} \left[e^{-\eta \ell(a)} \right]$$

- $\ell(p_\theta, y) = -\ln p_\theta(y)$ is *1-mixable* under $\zeta(p') = \mathbb{E}_{\theta \sim p'} p_\theta$ since $\ell(p') = -\ln \mathbb{E}_{\theta \sim p'} p_\theta$

Definition (α-Exp-concavity)

A convex function f is called α-*exp-concave* if the mapping $x \mapsto e^{-\alpha f(x)}$ is a concave function

- η-Mixability is just η-Exp-concavity with $\zeta(p) := \mathbb{E}_{a \sim p}[a]$!

Example: Least Squares Estimation

- For $y, a \in \mathbb{R}^d$, $\ell(a, y) = \|a - y\|_2^2$ is η-exp-concave with $\eta = \frac{1}{4Y^2}$

$$R_T(\theta) = \frac{1}{2} \sum_{t=1}^{T} \|a_t - y_t\|_2^2 - \|\theta - y_t\|_2^2$$

$$\leqslant 2Y^2 \sum_{t=1}^{T} -\ln \mathbb{E}_{a \sim p_t} \left[e^{-\frac{1}{4Y^2}\|a_t - y_t\|_2^2} \right] + \ln e^{-\frac{1}{4Y^2}\|\theta - y_t\|_2^2}$$

$$= 2Y^2 \sum_{t=1}^{T} -\ln \mathbb{E}_{a \sim p_t} \left[\frac{e^{-\frac{1}{2\sigma^2}\|a_t - y_t\|_2^2}}{(2\pi\sigma^2)^{d/2}} \right] + \ln \frac{e^{-\frac{1}{2\sigma^2}\|\theta - y_t\|_2^2}}{(2\pi\sigma^2)^{d/2}} \quad (\sigma = \sqrt{2}Y)$$

$$= 2Y^2 \sum_{t=1}^{T} \ell_{\log}(p_t(y_t)) - \ell_{\log}(p_\theta(y_t))$$

- Predictions using single actions easier than with mixtures up to a range-dependent constant!

- For all *squared* losses, exp-concavity ranges **depend on domains $\mathcal{Y}_t$**

Lemma (van der Hoeven et al. '18)

For $t - 1, \ldots, T$, suppose the loss ℓ is η_t-mixable on $(\mathcal{A}, \mathcal{Y}_t)$ with $\mathcal{Y}_t \subseteq \mathcal{Y}$ for sub-fun ζ_t. Then the exponentially-weighted forecaster algorithm with nonincreasing learning rates $\eta_1 \geqslant \ldots \geqslant \eta_T > 0$ and substitution functions $\zeta_1, \ldots, \zeta_T$ achieves

$$\sum_{t=1}^{T} \ell(a_t, y_t) \leqslant \mathbb{E}_{\theta \sim \gamma} \left[\sum_{t=1}^{T} \ell(f_\theta(x_t), y_t) \right] + \frac{\mathrm{KL}(\gamma|\pi)}{\eta_T}$$

For all priors π and gamma such that $\mathrm{KL}(\gamma|\pi_t) < \infty$, provided that the knowlege $y_t \in \mathcal{Y}_t$ is correct.

- As it turns out, the cost for not knowing $\mathcal{Y}_t$ one-step in advance is $\mathcal{O}(\frac{1}{\eta_T})$ for squared losses!
- Aggregate any hyperparameter α on an exponentially spaced grid

$$R_T(\texttt{Aggregated}, \theta) \lesssim R_T(\alpha^\star, \theta) + \frac{\log \log \alpha^\star}{\eta}$$

Online Multiclass Logistic Regression

- $y_t \in \{1, \dots, K\}$, Actions: probabilities over K classes
- Log loss: $\ell(p, y) = -\ln p(y)$
- Comparators parameterized by matrix $\theta \in \mathbb{R}^{K \times d}$ as $p_{\theta,t}(y) \propto e^{(\theta x_t)_y}$

Non-adaptive Result: [Foster et al. '18]

Known $U \geqslant \|\theta\|$, unknown $X = \max_{t \in [T]} \|x_t\|$

$$R_T(\theta) \leqslant 5dK \ln\left(\frac{UXT}{dK} + e\right)$$

Adaptive Result:

We show, with both U, X unknown:

$$R_T(\theta) \leqslant \underbrace{5dK \ln\left(\frac{2\|\theta\|XT}{dK} + 2e\right)}_{\text{Adaptive rate}} + \underbrace{\mathcal{O}\left(\log \log T\right)}_{\text{Cost of adaptation}}$$

Aggregate $U \in \{2^i \varepsilon / \|x_1\| : i \in \mathbb{N}\}$: poor dependence on $\varepsilon X / \|x_1\|$
Aggregate again $\varepsilon \in \{2^{-i}\}$ to improve to $+\mathcal{O}(\log \log (X / \|x_1\|))$

Logistic Regression II: Efficient Algorithm

Non-adaptive Result: [Agarwal et al. '21]

Slightly worse rate but practical runtime:

$$R_T(\theta) = \widetilde{\mathcal{O}}\big(UXdK \ln T\big) \quad \text{in} \quad \widetilde{\mathcal{O}}\big(d^2K^3 + UXK^2\big) \quad \text{time/round}$$

Linear dependence on $\|\theta\| \to$ more to gain through adaptation

Adaptive Result:

We show, for any $\beta > 0$ with $\|\theta\|X \leqslant T^\beta$:

$$R_T(\theta) = \widetilde{\mathcal{O}}\big(\|\theta\|XdK \ln T\big) \quad \text{in} \quad \widetilde{\mathcal{O}}\big(d^2K^3 + T^\beta K^2\big) \quad \text{time/round}$$

Challenge: Keeping Runtime Low

- Aggregate over a finite grid of U + doubling trick on X
- Total runtime is dominated by slowest algorithm

Online Least-squares Estimation

- $y_t, a_t \in \mathbb{R}^d$, square loss $\ell(a,y) = \|a-y\|^2/2$
- $f_\theta = \theta \in \mathbb{R}^d$; $Y = \max \|y_t\|$

Non-adaptive result:

Gradient Descent tuned with Y and U, for $\|\theta\| \leqslant U$,

$$R_T(\theta) \leqslant 2Y^2 \ln\left(1 + \frac{U^2 T}{Y^2}\right) + \frac{Y^2}{2}$$

Adaptive result:

We show, for any $\theta \in \mathbb{R}^d$

$$R_T(\theta) \leqslant 2Y^2 \ln\left(2 + \frac{\|\theta\|^2 T}{Y^2}\right) + \mathcal{O}\left(Y^2 \log\log\left(\frac{Y^2}{\|\theta\|^2}\right)\right)$$

Challenge: Mixability depends on unknown range of y_t

- Clip to previous largest $\|y_s\|$ for $+Y^2$ cost

Online Linear Least-squares Regression

- $a_t, y_t \in \mathbb{R}$, features $x_t \in \mathbb{R}^d$, square loss $\ell(a, y) = |a - y|^2/2$
- $f_\theta(x_t) = \theta^\intercal x_t$; $\quad Y = \max \|y_t\|$ and $X = \max \|x_t\|$

Non-adaptive: [Vovk'01, Azoury-Warmuth'01]

VAW forecaster tuned with Y, X and $U \geqslant \|\theta\|$

$$R_T(\theta) \leqslant \frac{dY^2}{2} \ln\left(1 + \frac{U^2 X^2 T}{d^2 Y^2}\right) + \mathcal{O}(1)$$

Adaptive:

We show for any $\theta \in \mathbb{R}^d$,

$$R_T(\theta) \leqslant \frac{dY^2}{2} \ln\left(1 + \frac{\|\theta\|^2 X^2 T}{d^2 Y^2}\right) + \mathcal{O}\left(\log\left|\log\left(\frac{Y^2}{\|\theta\|^2 X^2}\right)\right|\right)$$

- Aggregate over regularization + clipping to maintain mixability
- Scale-invariance by setting the grid according to scale $\|x_1\|$

Conclusion

No cost for adaptation in many online learning tasks

- Logistic regression, least-squares estimation, least-squares regression

More results in paper:

- Normal location, nonparametric classes
- Matching lower bounds with dependence on U, Y, X

Thanks for your attention!