Machine Learning Theory. Lecture 11.

Wouter M. Koolen

- OCO with exp-concavity:
  - Regression and Portfolio optimisation problem motivation.
  - Exp-concavity.
  - Derivation of Online Newton Step algorithm.
  - Analysis
- Mixability
  - Prediction with Expert Advice Protocol
  - How to actually combine predictions?

Exp-Concavity

Three popular losses

- Square loss for regression ($y_t \in \mathbb{R}$)

$$\boldsymbol{u} \;\mapsto\; \left(\langle \boldsymbol{u}, \boldsymbol{x}_t \rangle - y_t\right)^2$$

- Logistic loss for classification ($y_t \in \{\pm 1\}$)

$$\boldsymbol{u} \;\mapsto\; \ln(1 + e^{-y_t \langle \boldsymbol{u}, \boldsymbol{x}_t \rangle})$$

- Logarithmic loss for portfolio optimisation

$$\boldsymbol{u} \;\mapsto\; -\ln\langle \boldsymbol{u}, \boldsymbol{x}_t \rangle$$

Convex but *not* strongly convex. Q: Doomed to $\sqrt{T}$ regret?

Exp-Concavity

Normal convexity:

$$f(\boldsymbol{w}) - f(\boldsymbol{u}) \leq \langle \boldsymbol{w} - \boldsymbol{u}, \nabla f(\boldsymbol{w}) \rangle$$

Function f is α-exp-concave if $e^{-\alpha f(\boldsymbol{u})}$ is concave. So

$$e^{-\alpha f(\boldsymbol{u})} - e^{-\alpha f(\boldsymbol{w})} \leq \langle \boldsymbol{u} - \boldsymbol{w}, -\alpha e^{-\alpha f(\boldsymbol{w})} \nabla f(\boldsymbol{w}) \rangle$$

that is

$$f(\boldsymbol{w}) - f(\boldsymbol{u}) \leq \frac{1}{\alpha} \ln \left(1 + \alpha \langle \boldsymbol{w} - \boldsymbol{u}, \nabla f(\boldsymbol{w}) \rangle\right)$$

And if we pick $2\gamma \leq \alpha$ such that $|2\gamma \langle \boldsymbol{w} - \boldsymbol{u}, \nabla f(\boldsymbol{w}) \rangle| \leq \frac{1}{4}$, we find

$$f(\boldsymbol{w}) - f(\boldsymbol{u}) \leq \langle \boldsymbol{w} - \boldsymbol{u}, \nabla f(\boldsymbol{w}) \rangle - \frac{\gamma}{2} \langle \boldsymbol{w} - \boldsymbol{u}, \nabla f(\boldsymbol{w}) \rangle^2$$

Motivation of Algorithm

We have

$$f(\boldsymbol{w}) - f(\boldsymbol{u}) \;\leq\; \langle \boldsymbol{w} - \boldsymbol{u}, \nabla f(\boldsymbol{w}) \rangle - c \langle \boldsymbol{w} - \boldsymbol{u}, \nabla f(\boldsymbol{w}) \rangle^2$$

Idea: run exponentially weighted average forecaster on the loss

$$\boldsymbol{u} \;\mapsto\; \langle \boldsymbol{u} - \boldsymbol{w}, \nabla f(\boldsymbol{w}) \rangle + c \langle \boldsymbol{w} - \boldsymbol{u}, \nabla f(\boldsymbol{w}) \rangle^2$$

Quadratics get us Gaussians (mean $\boldsymbol{x}_1 = \boldsymbol{0}$, variance $\boldsymbol{I}/\epsilon$).

Motivation of Algorithm

Maintain weights $P_t = \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ on $\mathbb{R}^d$ s.t. $\boldsymbol{\mu}_t \in \mathcal{U}$. Start with $\boldsymbol{\mu}_1 = \mathbf{0}, \boldsymbol{\Sigma}_1 = \boldsymbol{I}/\epsilon$.

1. Play $\boldsymbol{\mu}_t$. Incur $f_t(\boldsymbol{\mu}_t)$.
2. See gradient $\nabla_t = \nabla f_t(\boldsymbol{\mu}_t)$. Update weights (with learning rate $\frac{1}{c}$)

$$\tilde{P}_{t+1}(\boldsymbol{u}) = \frac{P_t(\boldsymbol{u}) e^{-\frac{1}{c}\left(-\langle \boldsymbol{u}-\boldsymbol{\mu}_t, \nabla_t\rangle + c\langle \boldsymbol{\mu}_t - \boldsymbol{u}, \nabla_t\rangle^2\right)}}{\text{normalisation}} = \mathcal{N}\left(\tilde{\boldsymbol{\mu}}_{t+1}, \boldsymbol{\Sigma}_{t+1}\right)$$

where $\tilde{\boldsymbol{\mu}}_{t+1} = \boldsymbol{\mu}_t - \frac{1}{c}\boldsymbol{\Sigma}_{t+1}\nabla_t$ and $\boldsymbol{\Sigma}_{t+1}^{-1} = \boldsymbol{\Sigma}_t^{-1} + \nabla_t \nabla_t^\intercal$.

3. Project (if necessary) to ensure mean in $\mathcal{U}$:

$$P_{t+1} = \operatorname*{arg\,min}_{P:\mathbb{E}_{\boldsymbol{u}\sim P}[\boldsymbol{u}]\in\mathcal{U}} \mathrm{KL}(P\|\tilde{P}_{t+1}) = \mathcal{N}\left(\boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1}\right),$$

where $\boldsymbol{\mu}_{t+1} = \operatorname{arg\,min}_{\boldsymbol{u}\in\mathcal{U}} \left(\boldsymbol{u} - \tilde{\boldsymbol{\mu}}_{t+1}\right)^\intercal \boldsymbol{\Sigma}_{t+1}^{-1} \left(\boldsymbol{u} - \tilde{\boldsymbol{\mu}}_{t+1}\right)$.

Analysis of Algorithm

See Hazan, Chapter 4.

Notation here	Notation Hazan
$\boldsymbol{\mu}_t$	x_t
$\boldsymbol{\Sigma}_t$	A_{t-1}^{-1}
c	2γ

Prediction with Expert Advice

Prediction with Expert Advice

Prediction with *expert advice* for loss function $\mathcal{L} : \mathcal{A} \times \mathcal{X} \to \mathbb{R}$:

Protocol: for $t = 1, 2, \ldots$

- Experts announce actions $a_t^1, \ldots, a_t^K \in \mathcal{A}$.
- Learner chooses an action $a_t \in \mathcal{A}$.
- Adversary reveals outcome $x_t \in \mathcal{X}$.
- Learner incurs loss $\mathcal{L}(a_t, x_t)$.

Goal is small regret w.r.t. best expert:

$$R_T = \sum_{t=1}^{T} \mathcal{L}(a_t, x_t) - \min_k \sum_{t=1}^{T} \mathcal{L}(a_t^k, x_t)$$

Hedge (ignores advice) gives $R_T \leq O(\sqrt{T})$. Can get $\ln K$ for mix loss.
Q: When can we do better?

Mixable loss: Reduction to mix loss

Crux: exp-concavity is convenient but *too strong*.

Definition: We say $\mathcal{L}(a, x)$ is η-*mixable* if

$$\forall P \in \triangle_{\mathcal{A}} \exists a_P \in \mathcal{A} \forall x \in \mathcal{X} \quad \mathcal{L}(a_P, x) \leq \frac{-1}{\eta} \ln \left(\mathop{\mathbb{E}}_{a \sim P} e^{-\eta \mathcal{L}(a,x)} \right)$$

Mapping from P to witness a_P called *substitution function*.

Sufficient to check mixability for all *binary support* P.

Note: mixability is reparametrisation invariant (exp-concavity not so).

Exp-concavity: mixable with the mean as the substitution function.

Mixable losses regret bound

Mixable losses behave just enough like the mix loss to carry the AA regret bound through.

Theorem 1. *For any η-mixable loss $\mathcal{L}$ there is an algorithm guaranteeing*

$$R_T \leq \frac{\ln K}{\eta}$$

Proof sketch. Run the AA with losses $\ell_t^k = \eta \mathcal{L}(a_t^k, x_t)$. Given weights $\boldsymbol{w}_t$, construct distribution P_t with $P_t(a) = \sum_{k:a_t^k = a} w_t^k$. Then play $a_t = a_{P_t}$. Then apply the bound for AA, obtaining a $\ln K/\eta$ mix loss regret bound. The actual loss incurred is smaller (by mixability). □

Interestingly, Vovk (JCSS 1998) shows the converse.

Square loss is mixable

Let's consider

$$\mathcal{L}(a, x) = (a - x)^2$$

where $\mathcal{A} = \mathcal{X} = [-1, +1]$.

Square loss is mixable with $\eta = \frac{1}{2}$ (exp-concave only with $\eta = 1/8$).

The substitution function is

$$\boldsymbol{w}, a^1, \ldots, a^K \;\mapsto\; \frac{m_{\frac{1}{2}}(-1) - m_{\frac{1}{2}}(+1)}{4}$$

where $m_\eta(x) = \frac{-1}{\eta} \ln \sum_{k=1}^K w^k e^{-\eta(a^k - x)^2}$

See (Vovk 1990, Haussler, Kivinen, Warmuth, 1998)

For $\mathcal{A} = \mathcal{X} = [-Y, +Y]$ the constant becomes $\eta = \frac{1}{2Y^2}$.

Mixable loss list

Popular mixable losses:

- mix loss, log loss, entropic loss
- square loss, Brier loss
- Hellinger loss $\mathcal{A} = \mathcal{X} = [0, 1]$:

$$\mathcal{L}(a, x) \;:=\; \frac{1}{2}\left((\sqrt{1-x} - \sqrt{1-a})^2 + (\sqrt{x} - \sqrt{a})\right)$$

Characterisation of mixability: (Van Erven, Reid, Williamson 2012).