Machine Learning Theory. Lecture 12.

Wouter M. Koolen

- OCO with exp-concavity:
  - Regression and Portfolio optimisation problem motivation.
  - Exp-concavity.
  - 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

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

Quadratic upper bound

Let's make the upper bound more interpretable. Roughly,
$\ln(1+x) \approx x - \frac{1}{2}x^2$.

To make it an upper bound, we can use $\ln(1+x) \le x - x^2/4$ for $|x| \le 1$.

Applying exp-concavity with $2\gamma = \min\left\{\alpha, \frac{1}{4GD}\right\}$, we find

$$f(\boldsymbol{w}) - f(\boldsymbol{u}) \;\le\; \underbrace{\langle \boldsymbol{w} - \boldsymbol{u}, \nabla f(\boldsymbol{w})\rangle}_{\text{tangent}} \;\underbrace{- \frac{\gamma}{2}\langle \boldsymbol{w} - \boldsymbol{u}, \nabla f(\boldsymbol{w})\rangle^2}_{\text{quadratic bonus}} \qquad (1)$$

This will be the starting point of the algorithm design.

ONS algorithm

The Online Newton Step algorithm starts from $\boldsymbol{x}_1 = \mathbf{0} \in \mathcal{U}$ and $\boldsymbol{A}_0 = \epsilon \boldsymbol{I}$, and updates as

$$\begin{aligned} \boldsymbol{A}_t &= \boldsymbol{A}_{t-1} + \nabla_t \nabla_t^\intercal \\ \boldsymbol{x}_{t+1} &= \Pi_{\mathcal{U}}^{\boldsymbol{A}_t} \left(\boldsymbol{x}_t - \frac{1}{\gamma} \boldsymbol{A}_t^{-1} \nabla_t \right) \end{aligned}$$

where $\Pi_{\mathcal{U}}^{\boldsymbol{A}_t}$ is the projection in $\|\cdot\|_{\boldsymbol{A}_t}^2$, i.e.
$\Pi_{\mathcal{U}}^{\boldsymbol{A}_t}(\boldsymbol{u}) = \arg\min_{\boldsymbol{x} \in \mathcal{U}} \ (\boldsymbol{x} - \boldsymbol{u})^\intercal \boldsymbol{A}_t (\boldsymbol{x} - \boldsymbol{u})$.

Note the timing: $\boldsymbol{A}_t$ and $\boldsymbol{x}_{t+1}$ are both based on t rounds.

ONS result

Lemma 1. *For losses satisfying* (1)*, ONS guarantees*

$$R_T \;\leq\; \frac{\gamma}{2}\epsilon D^2 + \frac{d}{2\gamma}\ln\left(1+\frac{TG^2}{\epsilon d}\right).$$

Tuning $\epsilon = \frac{d}{\gamma^2 D^2}$ (which is optimal for $T \to \infty$) gives

$$R_T \;\leq\; \frac{d}{2\gamma}\left(1+\ln\left(1+T\frac{\gamma^2 D^2 G^2}{d^2}\right)\right).$$

ONS result

Theorem 2. *For α-exp-concave losses, using $\gamma = \frac{1}{2}\min\left\{\alpha, \frac{1}{4GD}\right\}$, so $\frac{1}{2\gamma} = \max\left\{\frac{1}{\alpha}, 4GD\right\}$, ONS guarantees*

$$R_T \;\leq\; \max\left\{\frac{1}{\alpha}, 4GD\right\} d\left(1 + \ln\left(1 + \frac{T}{64d^2}\right)\right).$$

ONS analysis

We look at the distance of the iterates to optimality, in $\|\boldsymbol{x}\|_{\boldsymbol{A}_t}^2 = \boldsymbol{x}^\intercal \boldsymbol{A}_t \boldsymbol{x}$

$$
\begin{aligned}
&\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|_{\boldsymbol{A}_t}^2 \\
&\overset{\text{Pyth. Th}}{\leq} \left\| \boldsymbol{x}_t - \frac{1}{\gamma} \boldsymbol{A}_t^{-1} \nabla_t - \boldsymbol{x}^* \right\|_{\boldsymbol{A}_t}^2 \\
&= \|\boldsymbol{x}_t - \boldsymbol{x}^*\|_{\boldsymbol{A}_t}^2 - \frac{2}{\gamma} \langle \boldsymbol{x}_t - \boldsymbol{x}^*, \nabla_t \rangle + \frac{1}{\gamma^2} \nabla_t^\intercal \boldsymbol{A}_t^{-1} \nabla_t \\
&= \|\boldsymbol{x}_t - \boldsymbol{x}^*\|_{\boldsymbol{A}_{t-1}}^2 + \langle \boldsymbol{x}_t - \boldsymbol{x}^*, \nabla_t \rangle^2 - \frac{2}{\gamma} \langle \boldsymbol{x}_t - \boldsymbol{x}^*, \nabla_t \rangle + \frac{1}{\gamma^2} \nabla_t^\intercal \boldsymbol{A}_t^{-1} \nabla_t
\end{aligned}
$$

where the last line uses $\boldsymbol{A}_t = \boldsymbol{A}_{t-1} + \nabla_t \nabla_t^\intercal$.

Reorganising gives an upper bound on the right-hand-side of (1)

$$\begin{aligned}
&\langle \boldsymbol{x}_t - \boldsymbol{x}^*, \nabla_t \rangle - \frac{\gamma}{2} \langle \boldsymbol{x}_t - \boldsymbol{x}^*, \nabla_t \rangle^2 \\
&\leq \frac{\gamma}{2} \left(\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2_{\boldsymbol{A}_{t-1}} - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2_{\boldsymbol{A}_t} \right) + \frac{1}{2\gamma} \nabla_t^\intercal \boldsymbol{A}_t^{-1} \nabla_t .
\end{aligned}$$

As $\ln \det$ is concave and its derivative is the matrix inverse,

$$\nabla_t^\intercal \boldsymbol{A}_t^{-1} \nabla_t = \operatorname{tr}\left((\boldsymbol{A}_t - \boldsymbol{A}_{t-1}) \boldsymbol{A}_t^{-1} \right) \overset{\text{Tangent}}{\leq} \ln \det \boldsymbol{A}_t - \ln \det \boldsymbol{A}_{t-1}$$

Combination with (1) and telescoping over rounds gives

$$\sum_{t=1}^{T} \left(f_t(\boldsymbol{x}_t) - f_t(\boldsymbol{x}^*) \right) \leq \frac{\gamma}{2} \|\boldsymbol{x}^*\|^2_{\boldsymbol{A}_0} + \frac{1}{2\gamma} \left(\ln \det \boldsymbol{A}_T - \ln \det \boldsymbol{A}_0 \right) .$$

Eliminating the log-determinant

As $\operatorname{tr}\left(\nabla_t \nabla_t^\intercal\right) = \|\nabla_t\|^2 \leq G^2$, we have $\operatorname{tr}\left(\boldsymbol{A}_T\right) \leq d\epsilon + TG^2$. By concavity of $\ln\det$

$$\ln\det \boldsymbol{A}_T \;\leq\; d\ln\left(\epsilon + \frac{TG^2}{d}\right)$$

and using $\|\boldsymbol{x}^*\|^2 \leq D^2$, we conclude

$$R_T \;\leq\; \frac{\gamma}{2}\epsilon D^2 + \frac{d}{2\gamma}\ln\left(1 + \frac{TG^2}{\epsilon d}\right).$$

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 3. *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).