

Machine Learning Theory 2024

Lecture 11

Wouter M. Koolen

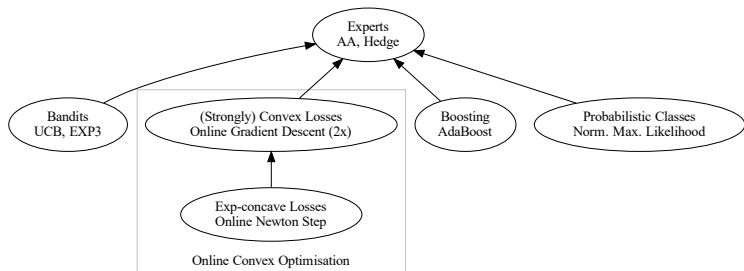
Download these slides now from elo.mastermath.nl!

- ▶ OCO with exp-concavity:
 - ▶ Regression and Portfolio optimisation problem motivation.
 - ▶ Exp-concavity.
 - ▶ Online Newton Step algorithm.
 - ▶ Analysis
 - ▶ Application: Concentration Inequality (Bonus)



Recap

Overview of Second Half of Course



Material: course notes on MLT website.

Recap: Online Convex Optimisation

General yet simple sequential decision problem.

Fix a convex set $\mathcal{U} \subseteq \mathbb{R}^d$.

Protocol

For $t = 1, 2, \dots$

- ▶ Learner chooses a point $\mathbf{w}_t \in \mathcal{U}$.
- ▶ Adversary reveals convex loss function $f_t : \mathcal{U} \rightarrow \mathbb{R}$.
- ▶ Learner's loss is $f_t(\mathbf{w}_t)$

Recap: Online Convex Optimisation

General yet simple sequential decision problem.

Fix a convex set $\mathcal{U} \subseteq \mathbb{R}^d$.

Protocol

For $t = 1, 2, \dots$

- ▶ Learner chooses a point $\mathbf{w}_t \in \mathcal{U}$.
- ▶ Adversary reveals convex loss function $f_t : \mathcal{U} \rightarrow \mathbb{R}$.
- ▶ Learner's loss is $f_t(\mathbf{w}_t)$

Objective:

Regret w.r.t. best point after T rounds:

$$R_T = \max_{\mathbf{u} \in \mathcal{U}} \sum_{t=1}^T (f_t(\mathbf{w}_t) - f_t(\mathbf{u}))$$

Recap: Results so far

We saw the Online Gradient Descent algorithm

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{U}}(\mathbf{w}_t - \eta_t \nabla f_t(\mathbf{w}_t))$$

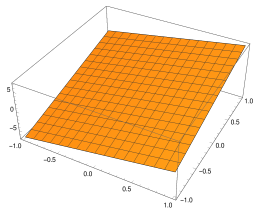
On Lipschitz **convex functions** OGD with $\eta \propto \frac{1}{\sqrt{T}}$ guarantees

$$R_T \leq GD\sqrt{T}.$$

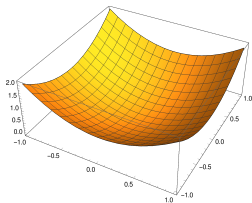
On **strongly convex functions** OGD with $\eta_t \propto \frac{1}{t}$ guarantees

$$R_T \leq O(\ln T)$$

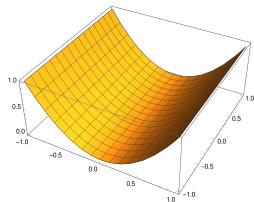
Where we are going today



Linear \subseteq Convex



Strongly Convex



Exp-concave

Exp-concavity

Exp-Concavity

Three popular losses

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

$$\mathbf{u} \mapsto (\langle \mathbf{u}, \mathbf{x}_t \rangle - y_t)^2$$

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

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

- ▶ **Logarithmic loss** for portfolio optimisation

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

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

Exp-Concavity

Normal convexity:

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

Strong convexity:

$$f(\mathbf{w}) - f(\mathbf{u}) \leq \langle \mathbf{w} - \mathbf{u}, \nabla f(\mathbf{w}) \rangle - \frac{\alpha}{2} \|\mathbf{w} - \mathbf{u}\|^2$$

Definition

A function $f : \mathcal{U} \rightarrow \mathbb{R}$ is called *exp-concave* to degree $\alpha \geq 0$ if $\mathbf{u} \mapsto e^{-\alpha f(\mathbf{u})}$ is concave.

Characterisations of Exp-Concavity I

In one dimension $\mathcal{U} \subseteq \mathbb{R}$, α -exp-concavity of f is equivalent to

$$f''(u) \geq \alpha(f'(u))^2$$

Characterisations of Exp-Concavity I

In one dimension $\mathcal{U} \subseteq \mathbb{R}$, α -exp-concavity of f is equivalent to

$$f''(u) \geq \alpha(f'(u))^2$$

This generalises:

Fact

A twice differentiable f is α -exp-concave at $\mathbf{u} \in \mathcal{U} \subseteq \mathbb{R}^d$ iff

$$\nabla^2 f(\mathbf{u}) \succeq \alpha \nabla f(\mathbf{u}) \nabla f(\mathbf{u})^\top. \quad (1)$$

Characterisations of Exp-Concavity II

Corollary

If f is α -exp concave for $\alpha > 0$ then

$$f(\mathbf{w}) - f(\mathbf{u}) \leq \frac{1}{\alpha} \ln(1 + \alpha \langle \mathbf{w} - \mathbf{u}, \nabla f(\mathbf{w}) \rangle) \quad \forall \mathbf{w}, \mathbf{u} \in \mathcal{U}. \quad (2)$$

Proof.

α -exp concavity implies

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

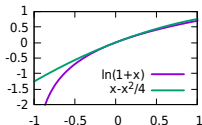
Multiply by $e^{\alpha f(\mathbf{w})}$, add 1, take \ln and divide by $\alpha > 0$. □

Towards a quadratic upper bound

By Taylor expansion in $x = 0$, $\ln(1 + x) \approx x - \frac{1}{2}x^2$.
Flips from upper to lower bound at $x = 0$.

Towards a quadratic upper bound

By Taylor expansion in $x = 0$, $\ln(1+x) \approx x - \frac{1}{2}x^2$.
Flips from upper to lower bound at $x = 0$.



Proposition

For $|x| \leq 1$ we have

$$\ln(1+x) \leq x - \frac{1}{4}x^2. \quad (3)$$

Proof.

Let's look at the gap $\ln(1+x) - x + x^2/4$. Its derivative, $\frac{1}{1+x} - 1 + \frac{x}{2}$ is zero when $x = 0$ or $x = 1$. The second derivative is $\frac{-1}{(1+x)^2} + \frac{1}{2}$, revealing that $x = 0$ is a maximum and $x = 1$ is a minimum. At $x = 0$ the gap is zero. So the gap is ≤ 0 for all $x \leq 1$. \square

Factor 2 alert!

Some sources use a **radius bound**

$$\|u\| \leq D \quad \forall u \in \mathcal{U},$$

while other sources use a **diameter bound**

$$\|u - w\| \leq D \quad \forall u, w \in \mathcal{U}.$$

By the triangle inequality, the diameter is at most twice the radius.

Following the previous lecture, these **slides** will use D to bound the **radius** of \mathcal{U} , while some other sources uses D for **diameter**. Be warned.

Quadratic upper bound

Lemma

Let $f : \mathcal{U} \rightarrow \mathbb{R}$ be α -exp-concave with bounded gradient $\|\nabla f(\mathbf{u})\| \leq G$ and radius $\|\mathbf{u}\| \leq D$ for all $\mathbf{u} \in \mathcal{U}$. Then for all $\gamma \leq \frac{1}{2} \min \left\{ \alpha, \frac{1}{2GD} \right\}$,

$$f(\mathbf{w}) - f(\mathbf{u}) \leq \underbrace{\langle \mathbf{w} - \mathbf{u}, \nabla f(\mathbf{w}) \rangle}_{\text{tangent}} - \underbrace{\frac{\gamma}{2} \langle \mathbf{w} - \mathbf{u}, \nabla f(\mathbf{w}) \rangle^2}_{\text{quadratic bonus}}. \quad (4)$$

Quadratic upper bound

Lemma

Let $f : \mathcal{U} \rightarrow \mathbb{R}$ be α -exp-concave with bounded gradient $\|\nabla f(\mathbf{u})\| \leq G$ and radius $\|\mathbf{u}\| \leq D$ for all $\mathbf{u} \in \mathcal{U}$. Then for all $\gamma \leq \frac{1}{2} \min \left\{ \alpha, \frac{1}{2GD} \right\}$,

$$f(\mathbf{w}) - f(\mathbf{u}) \leq \underbrace{\langle \mathbf{w} - \mathbf{u}, \nabla f(\mathbf{w}) \rangle}_{\text{tangent}} - \underbrace{\frac{\gamma}{2} \langle \mathbf{w} - \mathbf{u}, \nabla f(\mathbf{w}) \rangle^2}_{\text{quadratic bonus}}. \quad (4)$$

Proof.

(1) implies exp-concavity for degrees $\leq \alpha$. Applying (2) to $2\gamma \leq \alpha$ and then applying (3) using $|2\gamma \langle \mathbf{w} - \mathbf{u}, \nabla f(\mathbf{w}) \rangle| \leq \frac{\|\mathbf{w} - \mathbf{u}\| \|\nabla f(\mathbf{w})\|}{2GD} \leq 1$ give

$$\begin{aligned} f(\mathbf{w}) - f(\mathbf{u}) &\leq \frac{1}{2\gamma} \ln(1 + 2\gamma \langle \mathbf{w} - \mathbf{u}, \nabla f(\mathbf{w}) \rangle) \\ &\leq \frac{1}{2\gamma} \left(2\gamma \langle \mathbf{w} - \mathbf{u}, \nabla f(\mathbf{w}) \rangle - \frac{1}{4} (2\gamma \langle \mathbf{w} - \mathbf{u}, \nabla f(\mathbf{w}) \rangle)^2 \right) \\ &= \langle \mathbf{w} - \mathbf{u}, \nabla f(\mathbf{w}) \rangle - \frac{\gamma}{2} \langle \mathbf{w} - \mathbf{u}, \nabla f(\mathbf{w}) \rangle^2 \quad \square \end{aligned}$$

This will be the starting point of the algorithm design.

Online Newton Step

ONS algorithm

Let $\mathcal{U} \subseteq \mathbb{R}^d$ be a closed convex set containing $\mathbf{0}$.

The Online Newton Step (ONS) algorithm maintains an **iterate** $x_t \in \mathcal{U}$ and a positive definite $d \times d$ **matrix** $\mathbf{A}_{t-1} \succ \mathbf{0}$.

ONS algorithm

Let $\mathcal{U} \subseteq \mathbb{R}^d$ be a closed convex set containing $\mathbf{0}$.

The Online Newton Step (ONS) algorithm maintains an **iterate** $\mathbf{x}_t \in \mathcal{U}$ and a positive definite $d \times d$ **matrix** $\mathbf{A}_{t-1} \succ \mathbf{0}$.

Definition (Online Newton Step)

ONS with inverse learning rate $\epsilon > 0$ starts from

$$\mathbf{x}_1 = \mathbf{0} \in \mathcal{U} \quad \text{and} \quad \mathbf{A}_0 = \epsilon \mathbf{I}.$$

After receiving the gradient $\nabla_t := \nabla f_t(\mathbf{x}_t)$, it updates as

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{U}}^{\mathbf{A}_t} \left(\mathbf{x}_t - \frac{1}{\gamma} \mathbf{A}_t^{-1} \nabla_t \right) \quad \text{and} \quad \mathbf{A}_t = \mathbf{A}_{t-1} + \nabla_t \nabla_t^\top$$

where

$$\Pi_{\mathcal{U}}^{\mathbf{A}_t}(\mathbf{u}) = \arg \min_{\mathbf{x} \in \mathcal{U}} (\mathbf{x} - \mathbf{u})^\top \mathbf{A}_t (\mathbf{x} - \mathbf{u})$$

is the projection onto \mathcal{U} in the norm $\|\cdot\|_{\mathbf{A}_t}$.

Note the mixed timing: \mathbf{A}_t and \mathbf{x}_{t+1} are both based on t gradients.

ONS result

Theorem

For losses satisfying (4), ONS guarantees

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

ONS result

Theorem

For losses satisfying (4), ONS guarantees

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

Corollary

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

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

ONS result

Theorem

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

$$R_T \leq \max \left\{ \frac{1}{\alpha}, 2GD \right\} d \left(1 + \ln \left(1 + \frac{T}{16d^2} \right) \right)$$

ONS analysis I

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

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

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

Reorganising gives an upper bound on the right-hand-side of (4)

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

ONS analysis II

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

$$\nabla_t^T \mathbf{A}_t^{-1} \nabla_t = \text{tr}((\mathbf{A}_t - \mathbf{A}_{t-1}) \mathbf{A}_t^{-1}) \stackrel{\text{Tangent}}{\leq} \ln \det \mathbf{A}_t - \ln \det \mathbf{A}_{t-1}$$

Combination with (4) and telescoping over rounds gives

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

ONS analysis III

Recall that the **trace** is the sum of the eigenvalues, while the **log-determinant** is the sum of the logarithms of the eigenvalues.

As $\text{tr}(\nabla_t \nabla_t^T) = \|\nabla_t\|^2 \leq G^2$, we have $\text{tr}(\mathbf{A}_T) \leq d\epsilon + TG^2$. By concavity of the logarithm

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

Finally using $\|\mathbf{x}^*\|^2 \leq D^2$ and $\ln \det \mathbf{A}_0 = d \ln \epsilon$, we conclude

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

Application (not for exam)

Concentration from Online Learning

For i.i.d. zero-mean $Z_t \in [-1, +1]$ and λ_t predictable (function of $Z_1 \cdots Z_{t-1}$),

$$1 = \mathbb{E} \left[\prod_{t=1}^T (1 + \lambda_t Z_t) \right] = \mathbb{E} \left[e^{-\sum_{t=1}^T -\ln(1 + \lambda_t Z_t)} \right]$$

Concentration from Online Learning

For i.i.d. zero-mean $Z_t \in [-1, +1]$ and λ_t predictable (function of $Z_1 \cdots Z_{t-1}$),

$$1 = \mathbb{E} \left[\prod_{t=1}^T (1 + \lambda_t Z_t) \right] = \mathbb{E} \left[e^{-\sum_{t=1}^T -\ln(1 + \lambda_t Z_t)} \right]$$

So by Markov, for each $\delta \in (0, 1)$,

$$\delta \geq \mathbb{P} \left(e^{-\sum_{t=1}^T -\ln(1 + \lambda_t Z_t)} \geq \frac{1}{\delta} \right) = \mathbb{P} \left(\sum_{t=1}^T -\ln(1 + \lambda_t Z_t) \leq \ln \delta \right)$$

Concentration from Online Learning

For i.i.d. zero-mean $Z_t \in [-1, +1]$ and λ_t predictable (function of $Z_1 \cdots Z_{t-1}$),

$$1 = \mathbb{E} \left[\prod_{t=1}^T (1 + \lambda_t Z_t) \right] = \mathbb{E} \left[e^{-\sum_{t=1}^T -\ln(1 + \lambda_t Z_t)} \right]$$

So by Markov, for each $\delta \in (0, 1)$,

$$\delta \geq \mathbb{P} \left(e^{-\sum_{t=1}^T -\ln(1 + \lambda_t Z_t)} \geq \frac{1}{\delta} \right) = \mathbb{P} \left(\sum_{t=1}^T -\ln(1 + \lambda_t Z_t) \leq \ln \delta \right)$$

Letting λ_t be ONS iterates on 1d loss functions $\lambda \mapsto -\ln(1 + \lambda Z_t)$ gives

$$\sum_{t=1}^T -\ln(1 + \lambda_t Z_t) \leq \min_{\lambda} \sum_{t=1}^T -\ln(1 + \lambda Z_t) + O(\ln T)$$

Concentration from Online Learning

For i.i.d. zero-mean $Z_t \in [-1, +1]$ and λ_t predictable (function of $Z_1 \cdots Z_{t-1}$),

$$1 = \mathbb{E} \left[\prod_{t=1}^T (1 + \lambda_t Z_t) \right] = \mathbb{E} \left[e^{-\sum_{t=1}^T -\ln(1 + \lambda_t Z_t)} \right]$$

So by Markov, for each $\delta \in (0, 1)$,

$$\delta \geq \mathbb{P} \left(e^{-\sum_{t=1}^T -\ln(1 + \lambda_t Z_t)} \geq \frac{1}{\delta} \right) = \mathbb{P} \left(\sum_{t=1}^T -\ln(1 + \lambda_t Z_t) \leq \ln \delta \right)$$

Letting λ_t be ONS iterates on 1d loss functions $\lambda \mapsto -\ln(1 + \lambda Z_t)$ gives

$$\sum_{t=1}^T -\ln(1 + \lambda_t Z_t) \leq \min_{\lambda} \sum_{t=1}^T -\ln(1 + \lambda Z_t) + O(\ln T)$$

Further,

$$\min_{\lambda} \sum_{t=1}^T -\ln(1 + \lambda Z_t) \leq \min_{\lambda} \sum_{t=1}^T \left(-\lambda Z_t + \frac{1}{4}(\lambda Z_t)^2 \right) = -\frac{(\sum_{t=1}^T Z_t)^2}{\sum_{t=1}^T Z_t^2}$$

Concentration from Online Learning

For i.i.d. zero-mean $Z_t \in [-1, +1]$ and λ_t predictable (function of $Z_1 \cdots Z_{t-1}$),

$$1 = \mathbb{E} \left[\prod_{t=1}^T (1 + \lambda_t Z_t) \right] = \mathbb{E} \left[e^{-\sum_{t=1}^T -\ln(1 + \lambda_t Z_t)} \right]$$

So by Markov, for each $\delta \in (0, 1)$,

$$\delta \geq \mathbb{P} \left(e^{-\sum_{t=1}^T -\ln(1 + \lambda_t Z_t)} \geq \frac{1}{\delta} \right) = \mathbb{P} \left(\sum_{t=1}^T -\ln(1 + \lambda_t Z_t) \leq \ln \delta \right)$$

Letting λ_t be ONS iterates on 1d loss functions $\lambda \mapsto -\ln(1 + \lambda Z_t)$ gives

$$\sum_{t=1}^T -\ln(1 + \lambda_t Z_t) \leq \min_{\lambda} \sum_{t=1}^T -\ln(1 + \lambda Z_t) + O(\ln T)$$

Further,

$$\min_{\lambda} \sum_{t=1}^T -\ln(1 + \lambda Z_t) \leq \min_{\lambda} \sum_{t=1}^T \left(-\lambda Z_t + \frac{1}{4}(\lambda Z_t)^2 \right) = -\frac{(\sum_{t=1}^T Z_t)^2}{\sum_{t=1}^T Z_t^2}$$

All in all,

$$\mathbb{P} \left(\frac{(\sum_{t=1}^T Z_t)^2}{\sum_{t=1}^T Z_t^2} \geq \ln \frac{1}{\delta} + O(\ln T) \right) \leq \delta$$

Conclusion

Conclusion

Many practical losses are exp-concave. Assumption **between** convexity and **strong convexity**.

Learning algorithm ONS accumulates gradient directions into matrix.

$O(d \ln T)$ regret bound.

Unprojected update takes $O(d^2)$ time, projection often $O(d^3)$.