Machine Learning Theory 2023
Lecture 10

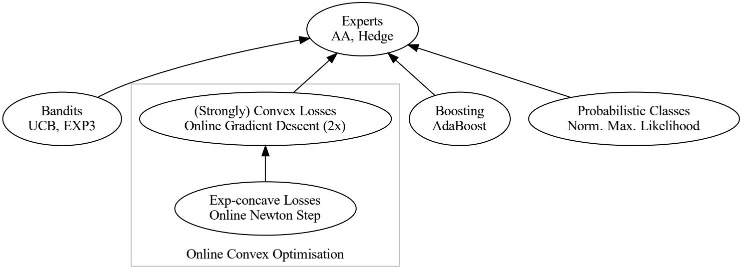
Wouter M. Koolen

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 $t = 1, 2, \ldots$

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

Recap: Online Convex Optimisation

General yet simple sequential decision problem.

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

Protocol

For $t = 1, 2, \ldots$

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

Objective:

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

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

Recap: Results so far

We saw the Online Gradient Descent algorithm

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

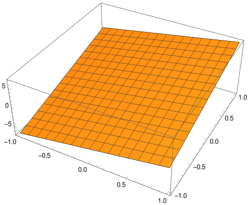
On Lipschitz **convex functions** OGD with $\eta \propto \frac{1}{\sqrt{T}}$ guarantees

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

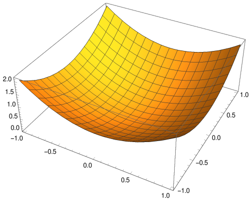
On **strongly convex functions** OGD with $\eta_t \propto \frac{1}{t}$ guarantees

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

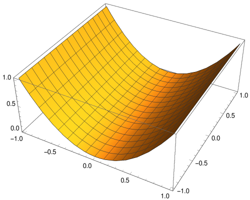
Where we are going today



Linear $\subseteq$ Convex



Stongly Convex



Exp-concave

Exp-concavity

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 ($y_t \in \{\pm 1\}$)

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

- **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:

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

Strong convexity:

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

Definition

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

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

Fact

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

$$\nabla^2 f(\boldsymbol{u}) \succeq \alpha \nabla f(\boldsymbol{u}) \nabla f(\boldsymbol{u})^\intercal. \tag{1}$$

Characterisations of Exp-Concavity II

Corollary

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

$$f(\boldsymbol{w}) - f(\boldsymbol{u}) \;\leq\; \frac{1}{\alpha}\ln\left(1 + \alpha\langle \boldsymbol{w} - \boldsymbol{u}, \nabla f(\boldsymbol{w})\rangle\right) \qquad \forall \boldsymbol{w}, \boldsymbol{u} \in \mathcal{U}. \tag{2}$$

Proof.

α-exp concavity implies

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

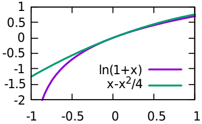
Multiply by $e^{\alpha f(\boldsymbol{w})}$, add 1, take ln and divide by $\alpha > 0$. □

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*

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

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 $x \leq 1$. □

Factor 2 alert!

Some sources use a **radius bound**

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

while other sources use a **diameter bound**

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

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

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

Quadratic upper bound

Lemma

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

$$f(\boldsymbol{w}) - f(\boldsymbol{u}) \;\le\; \underbrace{\langle \boldsymbol{w} - \boldsymbol{u}, \nabla f(\boldsymbol{w})\rangle}_{\text{tangent}} - \underbrace{\frac{\gamma}{2}\langle \boldsymbol{w} - \boldsymbol{u}, \nabla f(\boldsymbol{w})\rangle^2}_{\text{quadratic bonus}}. \tag{4}$$

Proof.

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

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

□

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** $\boldsymbol{x}_t \in \mathcal{U}$ and a positive definite $d \times d$ **matrix** $\boldsymbol{A}_{t-1} \succ \mathbf{0}$.

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** $\boldsymbol{x}_t \in \mathcal{U}$ and a positive definite $d \times d$ **matrix** $\boldsymbol{A}_{t-1} \succ \mathbf{0}$.

Definition (Online Newton Step)

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

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

After receiving the gradient $\nabla_t \coloneqq \nabla f_t(\boldsymbol{x}_t)$, it updates as

$$\boldsymbol{x}_{t+1} \;=\; \Pi_{\mathcal{U}}^{\boldsymbol{A}_t}\left(\boldsymbol{x}_t - \frac{1}{\gamma}\boldsymbol{A}_t^{-1}\nabla_t\right) \qquad \text{and} \qquad \boldsymbol{A}_t \;=\; \boldsymbol{A}_{t-1} + \nabla_t \nabla_t^\intercal$$

where

$$\Pi_{\mathcal{U}}^{\boldsymbol{A}_t}(\boldsymbol{u}) = \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathcal{U}} \; (\boldsymbol{x} - \boldsymbol{u})^\intercal \boldsymbol{A}_t (\boldsymbol{x} - \boldsymbol{u})$$

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

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

ONS result

Theorem

For losses satisfying (4), ONS guarantees

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

Corollary

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

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

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

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

ONS analysis I

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

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

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 (4)

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

ONS analysis II

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

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

Combination with (4) and telescoping over rounds gives

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

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 $\operatorname{tr}(\nabla_t \nabla_t^\intercal) = \|\nabla_t\|^2 \le G^2$, we have $\operatorname{tr}(\boldsymbol{A}_T) \le d\epsilon + TG^2$. By concavity of the logarithm

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

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

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

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,

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

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,

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

All in all,

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

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 $O(d^3)$.