

Machine Learning Theory. Lecture 11.

Wouter M. Koolen

- OCO with exp-concavity:
 - Regression and Portfolio optimisation problem motivation.
 - Exp-concavity.
 - Online Newton Step algorithm.
 - Analysis

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:

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

We say that a function f is α -*exp-concave* if $e^{-\alpha f(\mathbf{u})}$ is concave. So

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

that is

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

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 $\ln(1 + x) \leq x - x^2/4$ for $|x| \leq 1$.

Applying exp-concavity with $2\gamma = \min \left\{ \alpha, \frac{1}{4GD} \right\}$ (note: following the book, here we use D for diameter, not radius), we find

$$f(\mathbf{w}) - f(\mathbf{u}) \leq \underbrace{\langle \mathbf{w} - \mathbf{u}, \nabla f(\mathbf{w}) \rangle}_{\text{tangent}} - \underbrace{\frac{\gamma}{2} \langle \mathbf{w} - \mathbf{u}, \nabla f(\mathbf{w}) \rangle^2}_{\text{quadratic bonus}} \quad (1)$$

This will be the starting point of the algorithm design.

ONS algorithm

The Online Newton Step algorithm starts from $\mathbf{x}_1 = \mathbf{0} \in \mathcal{U}$ and $\mathbf{A}_0 = \epsilon \mathbf{I}$. After receiving the gradient $\nabla_t := \nabla f_t(\mathbf{x}_t)$, it updates as

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

where $\Pi_{\mathcal{U}}^{\mathbf{A}_t}$ is the projection in the norm $\|\cdot\|_{\mathbf{A}_t}^2$, i.e.

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

Note the mixed timing: \mathbf{A}_t and \mathbf{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 $T \rightarrow \infty$) gives

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

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 $\|\mathbf{x}\|_{\mathbf{A}_t}^2 = \mathbf{x}^\top \mathbf{A}_t \mathbf{x}$

$$\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_{\mathbf{A}_t}^2$$

$$\stackrel{\text{Pyth. Th}}{\leq} \left\| \mathbf{x}_t - \frac{1}{\gamma} \mathbf{A}_t^{-1} \nabla_t - \mathbf{x}^* \right\|_{\mathbf{A}_t}^2$$

$$= \|\mathbf{x}_t - \mathbf{x}^*\|_{\mathbf{A}_t}^2 - \frac{2}{\gamma} \langle \mathbf{x}_t - \mathbf{x}^*, \nabla_t \rangle + \frac{1}{\gamma^2} \nabla_t^\top \mathbf{A}_t^{-1} \nabla_t$$

$$= \|\mathbf{x}_t - \mathbf{x}^*\|_{\mathbf{A}_{t-1}}^2 + \langle \mathbf{x}_t - \mathbf{x}^*, \nabla_t \rangle^2 - \frac{2}{\gamma} \langle \mathbf{x}_t - \mathbf{x}^*, \nabla_t \rangle + \frac{1}{\gamma^2} \nabla_t^\top \mathbf{A}_t^{-1} \nabla_t$$

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 (1)

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

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

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

Combination with (1) and telescoping over rounds gives

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

Eliminating the log-determinant

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

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

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

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