Machine Learning Theory. Lecture 10.

Wouter M. Koolen

Online Convex Optimisation

- Gradient Descent for Convex Losses
- Online to Batch Conversion
- Gradient Descent for Strongly Convex Losses

Online Convex Optimisation

General yet simple sequential decision problem.

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

- Learner chooses a point $\boldsymbol{w}_t \in \mathcal{U}$.
- Adversary reveals 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:

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

Example loss functions

Setting	loss function $f_t(\boldsymbol{u})$
Hedge setting	$\boldsymbol{u}^\intercal \boldsymbol{\ell}_t$
Point prediction	$\|\boldsymbol{u} - \boldsymbol{x}_t\|^2$
Regression	$(\boldsymbol{u}^\intercal \boldsymbol{x}_t - y_t)^2$
Logistic regression	$\ln\left(1 + e^{-y_t \boldsymbol{u}^\intercal \boldsymbol{x}_t}\right)$
Hinge loss	$\max\{0, 1 - y_t \boldsymbol{u}^\intercal \boldsymbol{x}_t\}$
Investment	$-\ln(\boldsymbol{u}^\intercal \boldsymbol{x}_t)$
Offline optimization	$f(\boldsymbol{u})$

Online Gradient Descent (OGD)

Let $\mathcal{U}$ be a convex set containing $\mathbf{0}$. Fix $\eta > 0$. OGD plays

$$\boldsymbol{w}_1 = \mathbf{0} \qquad \text{and} \qquad \boldsymbol{w}_{t+1} = \Pi_{\mathcal{U}}\left(\boldsymbol{w}_t - \eta \nabla f_t(\boldsymbol{w}_t)\right)$$

where $\Pi_{\mathcal{U}}(\boldsymbol{w}) = \arg\min_{\boldsymbol{u}\in\mathcal{U}} \|\boldsymbol{u} - \boldsymbol{w}\|$ is the projection onto $\mathcal{U}$.

Theorem 1. *Let $\|\nabla f_t(\boldsymbol{u})\| \le G$ and $\|\boldsymbol{u}\| \le D$ for all $\boldsymbol{u} \in \mathcal{U}$. Then*

$$R_T = \max_{\boldsymbol{u}\in\mathcal{U}} \sum_{t=1}^{T} \left(f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{u})\right) \le \frac{1}{2\eta} D^2 + \frac{\eta}{2} T G^2$$

Corollary 2. *Tuning $\eta = \frac{D}{G\sqrt{T}}$ results in*

$$R_T^{\boldsymbol{u}} \le DG\sqrt{T}$$

Proof of GD regret bound

We have

$$f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{u}) \;\leq\; \langle \boldsymbol{w}_t - \boldsymbol{u}, \nabla f_t(w_t) \rangle$$

Moreover,

$$\begin{aligned}
\|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2 &= \|\Pi_{\mathcal{U}}\left(\boldsymbol{w}_t - \eta \nabla f_t(\boldsymbol{w}_t)\right) - \boldsymbol{u}\|^2 \\
&\leq \|\boldsymbol{w}_t - \eta \nabla f_t(\boldsymbol{w}_t) - \boldsymbol{u}\|^2 \\
&= \|\boldsymbol{w}_t - \boldsymbol{u}\|^2 - 2\eta \langle \boldsymbol{w}_t - \boldsymbol{u}, \nabla f_t(\boldsymbol{w}_t) \rangle + \eta^2 \|\nabla f_t(\boldsymbol{w}_t)\|^2
\end{aligned}$$

Hence

$$\langle \boldsymbol{w}_t - \boldsymbol{u}, \nabla f_t(\boldsymbol{w}_t) \rangle \;\leq\; \frac{\|\boldsymbol{w}_t - \boldsymbol{u}\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2}{2\eta} + \frac{\eta}{2} \|\nabla f_t(\boldsymbol{w}_t)\|^2$$

Summing over T rounds, we find

$$
\begin{aligned}
R_T^{\boldsymbol{u}} &\leq \sum_{t=1}^{T} \langle \boldsymbol{w}_t - \boldsymbol{u}, \nabla f_t(\boldsymbol{w}_t) \rangle \\
&\leq \underbrace{\sum_{t=1}^{T} \frac{\|\boldsymbol{w}_t - \boldsymbol{u}\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2}{2\eta}}_{\text{telescopes}} + \frac{\eta}{2} \sum_{t=1}^{T} \|\nabla f_t(\boldsymbol{w}_t)\|^2 \\
&\leq \frac{\|\boldsymbol{u}\|^2 - \cancel{\|\boldsymbol{w}_{T+1} - \boldsymbol{u}\|^2}}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\nabla f_t(\boldsymbol{w}_t)\|^2 \\
&\leq \frac{D^2}{2\eta} + \frac{\eta}{2} T G^2
\end{aligned}
$$

Online to Batch Conversion

Goal: obtain an estimator $\hat{\boldsymbol{w}}_T$ with small expected excess risk.

$$\mathop{\mathbb{E}}_{f_1,\ldots,f_T}\left[\mathop{\mathbb{E}}_{f}\left[f(\hat{\boldsymbol{w}}_T) - f(\boldsymbol{u}^*)\right]\right] \;\leq\; \text{small}$$

where the training set $f_1, \ldots, f_T$ and the test sample f are drawn i.i.d. and $\boldsymbol{u}^*$ optimises the risk $\boldsymbol{u} \mapsto \mathbb{E}_f[f(\boldsymbol{u})]$.

Idea: use online learning algorithm. Given training sample $f_1, \ldots, f_T$, the algorithm picks $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_T$. Let us define the *average iterate estimator*

$$\hat{\boldsymbol{w}}_T \;=\; \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{w}_t.$$

Theorem 3. *An online regret bound $R_T \leq B(T)$ implies*

$$\mathop{\mathbb{E}}_{\textit{iid } f_1,\ldots,f_T,\,f}\left[f\left(\hat{\boldsymbol{w}}_T\right) - f(\boldsymbol{u}^*)\right] \;\leq\; \frac{B(T)}{T}$$

Online to Batch Proof

$$\mathop{\mathbb{E}}_{\text{iid } f_1,\ldots,f_T,f}\left[f\left(\hat{\boldsymbol{w}}_T\right)-f(\boldsymbol{u}^*)\right]$$

$$\leq \mathop{\mathbb{E}}_{\text{iid } f_1,\ldots,f_T,f}\left[\frac{1}{T}\sum_{t=1}^{T}\left(f(\boldsymbol{w}_t)-f(\boldsymbol{u}^*)\right)\right]$$

$$= \mathop{\mathbb{E}}_{\text{iid } f_1,\ldots,f_T,f}\left[\frac{1}{T}\sum_{t=1}^{T}\left(f_t(\boldsymbol{w}_t)-f_t(\boldsymbol{u}^*)\right)\right] \leq \frac{B(T)}{T}$$

The first step is convexity of f. The last step uses that f and f_t have the same distribution (and $\boldsymbol{w}_t$ is not a function of f_t).

Strongly Convex Case

Convex function:

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

Strongly convex function:

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

Example: $f_t(\boldsymbol{w}) = \|\boldsymbol{w} - \boldsymbol{x}_t\|^2$.

Idea: could this extra knowledge help in the regret rate?

Online Gradient Descent with time-varying learning rate

$$\boldsymbol{w}_1 \;=\; \mathbf{0} \qquad \text{and} \qquad \boldsymbol{w}_{t+1} \;=\; \Pi_{\mathcal{U}}\left(\boldsymbol{w}_t - \eta_t \nabla f_t(\boldsymbol{w}_t)\right)$$

Theorem 4. *For strongly convex loss functions, OGD with learning rate* $\eta_t = \frac{1}{\alpha t}$ *ensures*

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

Proof

We start with

$$\begin{aligned}
\|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2 &= \|\Pi_{\mathcal{U}}\left(\boldsymbol{w}_t - \eta_t \nabla f_t(\boldsymbol{w}_t)\right) - \boldsymbol{u}\|^2 \\
&\leq \|\boldsymbol{w}_t - \eta_t \nabla f_t(\boldsymbol{w}_t) - \boldsymbol{u}\|^2 \\
&= \|\boldsymbol{w}_t - \boldsymbol{u}\|^2 - 2\eta_t \langle \boldsymbol{w}_t - \boldsymbol{u}, \nabla f_t(\boldsymbol{w}_t)\rangle + \eta_t^2 \|\nabla f_t(\boldsymbol{w}_t)\|^2
\end{aligned}$$

So that

$$\begin{aligned}
& f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{u}) \\
&\leq \langle \boldsymbol{w}_t - \boldsymbol{u}, \nabla f_t(\boldsymbol{w}_t)\rangle - \frac{\alpha}{2}\|\boldsymbol{w}_t - \boldsymbol{u}\|^2 \\
&\leq \frac{\|\boldsymbol{w}_t - \boldsymbol{u}\|^2 - \|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2 + \eta_t^2 \|\nabla f_t(\boldsymbol{w}_t)\|^2}{2\eta_t} - \frac{\alpha}{2}\|\boldsymbol{w}_t - \boldsymbol{u}\|^2 \\
&= \|\boldsymbol{w}_t - \boldsymbol{u}\|^2 \left(\frac{1}{2\eta_t} - \frac{\alpha}{2}\right) - \frac{\|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2}{2\eta_t} + \frac{\eta_t \|\nabla f_t(\boldsymbol{w}_t)\|^2}{2}
\end{aligned}$$

Key idea for telescoping:

$$\left(\frac{1}{\eta_{t+1}} - \alpha\right) \;=\; \frac{1}{\eta_t}$$

So

$$\eta_{t+1} \;=\; \frac{1}{\frac{1}{\eta_t} + \alpha}$$

A good starting point (cancelling the positive term after telescoping) is $\eta_1 \;=\; \frac{1}{\alpha}$. This leads to $\eta_t = \frac{1}{\alpha t}$. We then find

$$\begin{aligned}
R_T \;&=\; \sum_{t=1}^{T} f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{u}) \\
&\leq\; \sum_{t=1}^{T} \left(\|\boldsymbol{w}_t - \boldsymbol{u}\|^2 \left(\frac{1}{2\eta_t} - \frac{\alpha}{2}\right) - \frac{\|\boldsymbol{w}_{t+1} - \boldsymbol{u}\|^2}{2\eta_t} + \frac{\eta_t \|\nabla f_t(\boldsymbol{w}_t)\|^2}{2} \right) \\
&\leq\; \sum_{t=1}^{T} \frac{\|\nabla f_t(\boldsymbol{w}_t)\|^2}{2\alpha t} \;\leq\; \frac{G^2}{2\alpha}\left(1 + \ln T\right)
\end{aligned}$$