

Machine Learning Theory. Lecture 8.

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 $t = 1, 2, \dots$

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

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

Example loss functions

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

Online Gradient Descent (OGD)

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

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

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

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

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

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

$$R_T^u \leq DG\sqrt{T}$$

Proof of GD regret bound

We have

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

Moreover,

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

Hence

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

Summing over T rounds, we find

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

Online to Batch Conversion

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

$$\mathbb{E}_{f_1, \dots, f_T} \left[\mathbb{E}_f [f(\hat{\mathbf{w}}_T) - f(\mathbf{u}^*)] \right] \leq \text{small}$$

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

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

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

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

$$\mathbb{E}_{iid f_1, \dots, f_T, f} [f(\hat{\mathbf{w}}_T) - f(\mathbf{u}^*)] \leq \frac{B(T)}{T}$$

Online to Batch Proof

$$\begin{aligned} & \mathbb{E}_{\text{iid } f_1, \dots, f_T, f} [f(\hat{\mathbf{w}}_T) - f(\mathbf{u}^*)] \\ & \leq \mathbb{E}_{\text{iid } f_1, \dots, f_T, f} \left[\frac{1}{T} \sum_{t=1}^T (f(\mathbf{w}_t) - f(\mathbf{u}^*)) \right] \\ & = \mathbb{E}_{\text{iid } f_1, \dots, f_T, f} \left[\frac{1}{T} \sum_{t=1}^T (f_t(\mathbf{w}_t) - f_t(\mathbf{u}^*)) \right] \leq \frac{B(T)}{T} \end{aligned}$$

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

Strongly Convex Case

Convex function:

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

Strongly convex function:

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

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

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

Online Gradient Descent with time-varying learning rate

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

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

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

Proof

We start with

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

So that

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

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

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