Generalisation Bounds (5): Regret bounds for online learning

Qinfeng (Javen) Shi

The Australian Centre for Visual Technologies,
The University of Adelaide, Australia

2 Nov. 2012

Course Outline

Generalisation Bounds:

1. Basics
2. VC dimensions and bounds
3. Rademacher complexity and bounds
4. PAC Bayesian Bounds
5. Regret bounds for online learning (Today)
6. ...

Online Convex Optimisation

Online Convex Optimisation (OCO) can be seen as "an online player iteratively chooses a point from a non-empty, bounded, closed and convex set $\mathcal{C} \subset \mathbb{R}^n$"[1]

[1] Zinkevich03 and Hazan&Agarwal&Kale08 (Log regret algorithms for OCO)

OCO (2)

At iteration t, the algorithm $\mathcal{A}$ (the online player) chooses $\theta_t \in \mathcal{C}$. After committing to this choice, a convex cost function $f_t : \mathcal{C} \to \mathbb{R}$ is revealed (*i.e.* $f_t(\theta_t)$ is the cost).
That is (in general)

$$\theta_t = \mathcal{A}(\{f_1, \cdots, f_{t-1}\})$$

OCO (3)

Denote the number of iterations by T, the goal of OCO is to minimise the Regret

$$\text{Regret}(\mathcal{A}, \{f_1, \cdots, f_T\}) = \sum_{t=1}^{T} f_t(\theta_t) - \min_{\theta} \sum_{t=1}^{T} f_t(\theta). \quad (1)$$

Online Learning

Online Learning (OL)[2]:
At iteration t, the algorithm $\mathcal{A}$ receives a instance $x_t \in \mathbb{R}^n$ and is then required to predict the output [3] $\hat{y}_t = h(x_t; \theta_t)$. After predicting $\hat{y}_t$, the true output y_t is revealed and a loss $\ell(\theta_t; (x_t, y_t))$ occurs. Then $\ell(\theta; (x_t, y_t)) \rightarrow \theta_{t+1}$. Denote $\#$ iter. by T, the goal of OL is to minimise the Regret

$$\text{Regret}(\mathcal{A}, \{(x_1, y_1), \cdots, (x_T, y_T)\}) = \sum_{t=1}^{T} \ell(\theta_t; (x_t, y_t)) - \min_{\theta} \sum_{t=1}^{T} \ell(\theta; (x_t, y_t)). \tag{2}$$

View OL as an OCO: $\ell(\theta; (x_{t-1}, y_{t-1})) \rightarrow \theta_t$ is essentially picking θ_t in OCO. $\ell(\theta_t; (x_t, y_t))$ is the cost function $f_t(\theta_t)$.

---

[2] more general OL can be described without θ and h

[3] label in classification, or response in regression

Online Learning — Loss functions

The loss ℓ can be any loss function in Empirical Risk Minimisation (ERM).

Table 5: Scalar loss functions and their derivatives, depending on $f := \langle w, x \rangle$, and y.

	Loss $\tilde{l}(f, y)$	Derivative $\tilde{l}'(f, y)$
Hinge (Bennett and Mangasarian, 1992)	$\max(0, 1 - yf)$	0 if $yf \geq 1$ and $-y$ otherwise
Squared Hinge (Keerthi and DeCoste, 2005)	$\frac{1}{2}\max(0, 1 - yf)^2$	0 if $yf \geq 1$ and $f - y$ otherwise
Exponential (Cowell et al., 1999)	$\exp(-yf)$	$-y \exp(-yf)$
Logistic (Collins et al., 2000)	$\log(1 + \exp(-yf))$	$-y/(1 + \exp(-yf))$
Novelty (Schölkopf et al., 2001)	$\max(0, \rho - f)$	0 if $f \geq \rho$ and -1 otherwise
Least mean squares (Williams, 1998)	$\frac{1}{2}(f - y)^2$	$f - y$
Least absolute deviation	$\lvert f - y \rvert$	$\operatorname{sgn}(f - y)$
Quantile regression (Koenker, 2005)	$\max(\tau(f - y), (1 - \tau)(y - f))$	τ if $f > y$ and $\tau - 1$ otherwise
ϵ-insensitive (Vapnik et al., 1997)	$\max(0, \lvert f - y \rvert - \epsilon)$	0 if $\lvert f - y \rvert \leq \epsilon$, else $\operatorname{sgn}(f - y)$
Huber's robust loss (Müller et al., 1997)	$\frac{1}{2}(f - y)^2$ if $\lvert f - y \rvert \leq 1$, else $\lvert f - y \rvert - \frac{1}{2}$	$f - y$ if $\lvert f - y \rvert \leq 1$, else $\operatorname{sgn}(f - y)$
Poisson regression (Cressie, 1993)	$\exp(f) - yf$	$\exp(f) - y$

Table 6: Vectorial loss functions and their derivatives, depending on the vector $f := Wx$ and on y.

	Loss	Derivative
Soft-Margin Multiclass (Taskar et al., 2004) (Crammer and Singer, 2003)	$\max_{y'}(f_{y'} - f_y + \Delta(y, y'))$	$e_{y^*} - e_y$ where y^* is the argmax of the loss
Scaled Soft-Margin Multiclass (Tsochantaridis et al., 2005)	$\max_{y'} \Gamma(y, y')(f_{y'} - f_y + \Delta(y, y'))$	$\Gamma(y, y')(e_{y^*} - e_y)$ where y^* is the argmax of the loss
Softmax Multiclass (Cowell et al., 1999)	$\log \sum_{y'} \exp(f_{y'}) - f_y$	$\left[\sum_{y'} e_{y'} \exp(f'_y)\right] / \sum_{y'} \exp(f'_y) - e_y$
Multivariate Regression	$\frac{1}{2}(f - y)^\top M(f - y)$ where $M \succeq 0$	$M(f - y)$

Table 5 in TeoVisSmo09 bundle method for risk minimisation. Note f here is not the cost function f_t in OCO.

Typical regret bounds

For OCO algorithms, if the f_t is strongly-convex and differentiable (sometimes twice differentiable), we often have

$$\text{Regret}(\mathcal{A}, \{f_1, \cdots, f_T\}) \leq O(\log T).$$

Typical assumptions

Denote D the diameter of the underlying convex set $\mathcal{C}$. *i.e.*

$$D = \max_{\theta,\theta' \in \mathcal{C}} \|\theta - \theta'\|_2$$

Assume f_t

- differentiable (twice differentiable needed when the Hessian is used (*e.g.* Newton method))
- bounded gradient by G *i.e.*

$$\sup_{\theta \in \mathcal{C}, t \in [T]} \|\nabla f_t(\theta)\|_2 \leq G$$

- H-strongly convex

$$f_t(\theta) - f_t(\theta') \geq \nabla f_t(\theta')^T(\theta - \theta') + \frac{H}{2}\|\theta - \theta'\|_2^2$$

Online Gradient Descent

Input: Convex Set $\mathcal{C} \subset \mathbb{R}^n$, step sizes $\eta_1, \eta_2, \dots \geq 0$, initial $\theta_1 \in \mathcal{C}$.
In iteration 1, use θ_1.
In iteration $t > 1$: use

$$\theta_t = \Pi_{\mathcal{C}}(\theta_{t-1} - \eta_t \nabla f_{t-1}(\theta_{t-1})).$$

Here $\Pi_{\mathcal{C}}$ denotes the projection onto nearest point in $\mathcal{C}$, that is

$$\Pi_{\mathcal{C}}(\theta) = \operatorname*{argmin}_{\theta' \in \mathcal{C}} \|\theta - \theta'\|_2.$$

Regret bound for OGD

Let $\theta^* \in \mathrm{argmin}_{\theta \in \mathcal{C}} \sum_{t=1}^{T} f_t(\theta)$, recall regret def (*i.e.* (1)),

$$\text{Regret}_T(OGD) = \sum_{t=1}^{T} f_t(\theta_t) - \sum_{t=1}^{T} f_t(\theta^*).$$

Theorem (Regret on OGD)

For OGD with step sizes $\eta_t = \frac{1}{H(t-1)}, 2 \leq t \leq T$, for all $T \geq 2$,

$$\textit{Regret}_T(OGD) \leq \frac{G^2}{2H}(1 + \log T). \tag{3}$$

Regret bound for OGD — proof

f_t is H-strongly convex, we have

$$f(\theta^*) - f(\theta_t) \geq \nabla f_t(\theta_t)^T(\theta^* - \theta_t) + \frac{H}{2}\|\theta^* - \theta_t\|_2^2$$
$$\Rightarrow f(\theta_t) - f(\theta^*) \leq \nabla f_t(\theta_t)^T(\theta_t - \theta^*) - \frac{H}{2}\|\theta^* - \theta_t\|_2^2.$$

Claim:

$$\nabla f_t(\theta_t)^T(\theta_t - \theta^*) \leq \frac{\|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2}{2\eta_{t+1}} + \frac{\eta_{t+1} G^2}{2} \tag{4}$$

Regret bound for OGD — proof

$$f(\theta_t) - f(\theta^*) \le \frac{\|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2}{2\eta_{t+1}} + \frac{\eta_{t+1} G^2}{2} - \frac{H}{2}\|\theta^* - \theta_t\|_2^2. \tag{5}$$

Sum up (5) for $t = 1, \cdots, T$, we have

$$\begin{aligned}
\sum_{t=1}^{T}(f(\theta_t) - f(\theta^*)) &\le \frac{1}{2}\left(\frac{1}{\eta_2} - H\right)\|\theta_1 - \theta^*\|^2 - \frac{1}{2\eta_{T+1}}\|\theta_{T+1} - \theta^*\|^2 \\
&+ \frac{1}{2}\sum_{t=2}^{T}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - H\right)\|\theta_t - \theta^*\|^2 + \frac{G^2}{2}\sum_{t=1}^{T}\eta_{t+1} \\
&\le 0 + \frac{G^2}{2H}\sum_{t=1}^{T}\frac{1}{t} \qquad \left(\text{recall } \eta_t = \frac{1}{H(t-1)}, \text{ blue} = 0, \text{ red} \le 0\right) \\
&\le \frac{G^2}{2H}(1 + \log T).
\end{aligned}$$

Regret bound for OGD — proof

To prove the Claim:

$$\nabla f_t(\theta_t)^T(\theta_t - \theta^*) \leq \frac{\|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2}{2\eta_{t+1}} + \frac{\eta_{t+1} G^2}{2}$$

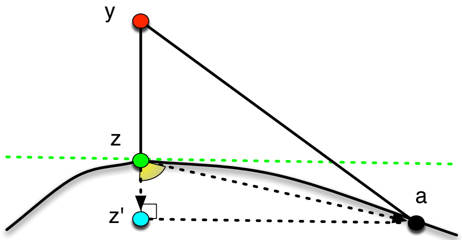
$$\begin{aligned}
&\|\theta_{t+1} - \theta^*\|^2 \\
&= \|\Pi_{\mathcal{C}}(\theta_t - \eta_{t+1}\nabla f_t(\theta_t)) - \theta^*\|^2 \\
&\leq \|(\theta_t - \eta_{t+1}\nabla f_t(\theta_t)) - \theta^*\|^2 \quad \text{(a property of proj onto a convex set)} \\
&= \|\theta_t - \theta^*\|^2 + \eta_{t+1}^2\|\nabla f_t(\theta_t))\|^2 - 2\eta_{t+1}\nabla f_t(\theta_t)^T(\theta_t - \theta^*) \\
&\leq \|\theta_t - \theta^*\|^2 + \eta_{t+1}^2 G^2 - 2\eta_{t+1}\nabla f_t(\theta_t)^T(\theta_t - \theta^*).
\end{aligned}$$

Rearrange the inequality and divide by $2\eta_{t+1}$ yields the claim.

Regret bound for OGD — proof

A property of projection onto a convex set:
Let $\mathcal{C} \subset \mathbb{R}^n$ be a convex set, $y \in \mathbb{R}^n$ and $z = \Pi_{\mathcal{C}} y$ be the projection of y onto $\mathcal{C}$. The for any point $a \in \mathcal{C}$,

$$\|y - a\|^2 \geq \|z - a\|^2.$$



Intuition: Convexity of $\mathcal{C} \Rightarrow (z - y)^T(a - z) \geq 0$ (*i.e.* yellow angle acute). $\Rightarrow \|y - a\|^2 \geq \|z - a\|^2$.
(See Lemma 8 of Hazan etal 08 for proof)

Non-necessary assumptions

Relax OGD assumptions on f_t to following

- non-differentiable (pick a good sub-gradient)
- bounded (sub)-gradient by G *i.e.*

$$\sup_{\theta \in \mathcal{C}, t \in [T]} \|\nabla f_t(\theta)\|_2 \leq G$$

- H-strongly convex (for (sub)-gradient)

$$f_t(\theta) - f_t(\theta') \geq \nabla f_t(\theta')^T (\theta - \theta') + \frac{H}{2} \|\theta - \theta'\|_2^2$$

Non-necessary projection step

In OGD, the projection step *i.e.*

$$\theta_t = \Pi_{\mathcal{C}}(\theta_{t-1} - \eta_t \nabla f_{t-1}(\theta_{t-1})),$$

may be removed. Projection is just to ensure every θ_t is still a feasible point. If this is not a problem, without projection, we still have

$$\begin{aligned}
&\|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2 \\
&= \|\theta_t - \theta^*\|^2 - \|\theta_t - \eta_t \nabla f_t(\eta_t) - \theta^*\|^2 \\
&= \|\theta_t - \theta^*\|^2 - \|(\theta_t - \theta^*) - \eta_t \nabla f_t(\eta_t)\|^2 \\
&= 2\eta_{t+1} \nabla f_t(\theta_t)^T(\theta_t - \theta^*) - \eta_{t+1}^2 (\nabla f_t(\eta_t))^2 \\
&\geq 2\eta_{t+1} \nabla f_t(\theta_t)^T(\theta_t - \theta^*) - \eta_{t+1}^2 G^2
\end{aligned}$$

Above still yields the claim

$$\nabla f_t(\theta_t)^T(\theta_t - \theta^*) \leq \frac{\|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2}{2\eta_{t+1}} + \frac{\eta_{t+1} G^2}{2}$$

Pegasos: Primal Estimated sub-GrAdient SOlver for SVM

Pegasos (Shalev-Shwartz&Singer&Srebro07 and Shalev-Shwartz&Singer&Srebro&Cotter09) can be seen as OGD with

$$f_t(\theta) = \frac{H}{2}\|\theta\|^2 + [1 - y_t \langle \theta, x_t \rangle]_+,$$

However $f_t(\theta)$ is not differentiable at where $1 - y_t \langle \theta, x_t \rangle = 0$, which violates the old assumptions of OGD.
Remedy: pick sub-Gradient $\nabla f_t(\theta_t) = 0$ where $1 - y_t \langle \theta, x_t \rangle = 0$ and let $\nabla f_t(\theta_t)$ be the gradient where differentiable. Now even when $1 - y_t \langle \theta, x_t \rangle = 0$, H-strongly convexity (from $\frac{H}{2}\|\theta\|^2$) and bounded (sub-)gradient still hold.

See Lemma 1 of Shalev-Shwartz&Singer&Srebro07 which gives the same regret bound as OGD.