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

$$\theta_t = \mathcal{A}(\{f_1, \cdots, f_{t-1}\})$$
Denote the number of iterations by $T$, the goal of OCO is to minimise the Regret

$$\text{Regret}(A, \{f_1, \cdots, f_T\}) = \sum_{t=1}^{T} f_t(\theta_t) - \min_{\theta} \sum_{t=1}^{T} f_t(\theta). \quad (1)$$
Online Learning (OL)^2:
At iteration \( t \), the algorithm \( \mathcal{A} \) receives an 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 \( \theta_t \) 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 \( \theta \) and \( h \)

\(^3\)label in classification, or response in regression
Online Learning — Loss functions

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

<table>
<thead>
<tr>
<th>Loss ( l(f, y) )</th>
<th>Derivative ( l'(f, y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hinge (Bennett and Mangasarian, 1992)</td>
<td>( \max(0, 1 - yf) )</td>
</tr>
<tr>
<td>Squared Hinge (Keerthi and DeCoste, 2005)</td>
<td>( \frac{1}{2} \max(0, 1 - yf)^2 )</td>
</tr>
<tr>
<td>Exponential (Cowell et al., 1999)</td>
<td>( \exp(-yf) )</td>
</tr>
<tr>
<td>Logistic (Collins et al., 2000)</td>
<td>( \log(1 + \exp(-yf)) )</td>
</tr>
<tr>
<td>Novelty (Schölkopf et al., 2001)</td>
<td>( \max(0, \rho - f) )</td>
</tr>
<tr>
<td>Least mean squares (Williams, 1998)</td>
<td>( \frac{1}{2}(f - y)^2 )</td>
</tr>
<tr>
<td>Least absolute deviation</td>
<td>(</td>
</tr>
<tr>
<td>Quantile regression (Koenker, 2005)</td>
<td>( \max(\tau(f - y), (1 - \tau)(y - f)) )</td>
</tr>
<tr>
<td>( \epsilon )-insensitive (Vapnik et al., 1997)</td>
<td>( \max(0,</td>
</tr>
<tr>
<td>Huber’s robust loss (Müller et al., 1997)</td>
<td>( \frac{1}{2}(f - y)^2 ) if (</td>
</tr>
<tr>
<td>Poisson regression (Cressie, 1993)</td>
<td>( \exp(f - yf) )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Loss</th>
<th>Derivative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soft-Margin Multiclass (Taskar et al., 2004) (Crammer and Singer, 2003)</td>
<td>( \max_y (f_{y'} - f_y + \Delta(y, y')) )</td>
</tr>
<tr>
<td>Scaled Soft-Margin Multiclass (Tsochantaridis et al., 2005)</td>
<td>( \max_y \Gamma(y, y') (f_{y'} - f_y + \Delta(y, y')) )</td>
</tr>
<tr>
<td>Softmax Multiclass (Cowell et al., 1999)</td>
<td>( \log \sum_{y'} \exp(f_{y'}) - f_y )</td>
</tr>
<tr>
<td>Multivariate Regression</td>
<td>( \frac{1}{2}(f - y)' M(f - y) ) where ( M \geq 0 )</td>
</tr>
</tbody>
</table>

Table 5 in TeoVisSmo09 bundle method for risk minimisation. Note \( f \) here is not the cost function \( f_t \) in OCO.
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 $C$. 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 $\eta_1, \eta_2, \cdots \geq 0$, initial $\theta_1 \in \mathcal{C}$.

In iteration 1, use $\theta_1$.

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

$$
\Pi_{\mathcal{C}}(\theta) = \arg\min_{\theta' \in \mathcal{C}} \|\theta - \theta'\|_2.
$$
Let $\theta^* \in \arg\min_{\theta \in \mathcal{C}} \sum_{t=1}^{T} f_t(\theta)$, 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 $T \geq 2$,*

$$\text{Regret}_T(OGD) \leq \frac{G^2}{2H}(1 + \log T).$$ (3)
$f_t$ is $H$-strongly convex, we have

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

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

Claim:

$$\nabla f_t(\theta_t)^T(\theta_t - \theta^*) \leq \frac{\|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2}{2\eta_{t+1}} + \frac{\eta_{t+1}G^2}{2} \quad (4)$$
\[ f(\theta_t) - f(\theta^*) \leq \frac{\|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2}{2\eta_{t+1}} + \frac{\eta_{t+1} G^2}{2} - \frac{H}{2} \|\theta^* - \theta_t\|^2. \]  

(5)

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

\[
\sum_{t=1}^{T} (f(\theta_t) - f(\theta^*)) \leq \frac{1}{2} \left( \frac{1}{\eta_2} - H \right) \|\theta_1 - \theta^*\|^2 - \frac{1}{2\eta_{T+1}} \|\theta_{T+1} - \theta^*\|^2 \\
+ \frac{1}{2} \sum_{t=2}^{T} \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - H \right) \|\theta_t - \theta^*\|^2 + \frac{G^2}{2} \sum_{t=1}^{T} \eta_{t+1} \\
\leq 0 + \frac{G^2}{2H} \sum_{t=1}^{T} \frac{1}{t} \quad \text{(recall } \eta_t = \frac{1}{H(t-1)}, \text{ blue } = 0, \text{ red } \leq 0) \\
\leq \frac{G^2}{2H} (1 + \log T). \]
To prove the Claim:

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

\[
\|\theta_{t+1} - \theta^*\|^2 \\
= \|\Pi_C(\theta_t - \eta_{t+1} \nabla f_t(\theta_t)) - \theta^*\|^2 \\
\leq \|(\theta_t - \eta_{t+1} \nabla f_t(\theta_t)) - \theta^*\|^2 \quad \text{(a property of proj onto a convex set)} \\
= \|\theta_t - \theta^*\|^2 + \eta_{t+1}^2 \|\nabla f_t(\theta_t)\|^2 - 2\eta_{t+1} \nabla f_t(\theta_t)^T (\theta_t - \theta^*) \\
\leq \|\theta_t - \theta^*\|^2 + \eta_{t+1}^2 G^2 - 2\eta_{t+1} \nabla f_t(\theta_t)^T (\theta_t - \theta^*).
\]

Rearrange the inequality and divide by $2\eta_{t+1}$ yields the claim.
A property of projection onto a convex set:
Let $C \subset \mathbb{R}^n$ be a convex set, $y \in \mathbb{R}^n$ and $z = \Pi_C y$ be the projection of $y$ onto $C$. The for any point $a \in C$,

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

Intuition: Convexity of $C \Rightarrow (z - y)^T (a - z) \geq 0$ (i.e. yellow angle acute). $\Rightarrow \|y - a\|^2 \geq \|z - a\|^2$.
(See Lemma 8 of Hazan etal 08 for proof)
Relax OGD assumptions on $f_t$ to following

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

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

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

$$f_t(\theta) - f_t(\theta') \geq \nabla f_t(\theta')^T (\theta - \theta') + \frac{H}{2} \| \theta - \theta' \|^2$$
In OGD, the projection step \( i.e. \theta_t = \Pi_C(\theta_{t-1} - \eta_t \nabla f_{t-1}(\theta_{t-1})) \),

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

\[
\|\theta_t - \theta^*\|^2 - \|\theta_{t+1} - \theta^*\|^2
= \|\theta_t - \theta^*\|^2 - \|\theta_t - \eta_t \nabla f_t(\eta_t) - \theta^*\|^2
= \|\theta_t - \theta^*\|^2 - \|\theta_t - \theta^* - \eta_t \nabla f_t(\eta_t)\|^2
= 2\eta_{t+1} \nabla f_t(\theta_t)^T (\theta_t - \theta^*) - \eta_{t+1}^2 (\nabla f_t(\eta_t))^2
\geq 2\eta_{t+1} \nabla f_t(\theta_t)^T (\theta_t - \theta^*) - \eta_{t+1}^2 G^2
\]

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 (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.