Lecture 14: Online Gradient Descent
Online Learning

• Problem specified by: $\ell : \overline{\mathcal{H}} \times \mathcal{Z} \rightarrow \mathbb{R}$

• We want to compete with hypothesis class $\mathcal{H} \subseteq \overline{\mathcal{H}}$

• Rule: $A : \mathcal{Z}^* \rightarrow \overline{\mathcal{H}}$ attains regret $Reg(m)$ on $\mathcal{H}$ if for any sequence:

$$\frac{1}{m} \sum_{t=1}^{m} \ell (A(z_1, \ldots, z_{t-1}), z_t) \leq \inf_{h \in \mathcal{H}} \frac{1}{m} \sum_{t=1}^{m} \ell (h, z_t) + Reg(m)$$

• For what problems and hypothesis classes can we get $Reg(m) \rightarrow 0$?

• What is the best regret we can attain for a hypothesis class?

• How and when are online regret and statistical learning related?

Emphasis on:

• Relationship to statistical (PAC) learning

• Computationally easy, incremental, learning rules
What’s Online Learnable?

• Finite cardinality classes
  • Realizable: \( \text{Reg}(m) \leq \frac{\log|\mathcal{H}|}{m} \), using Halving, matching statistical learning
  • Non-realizable case: ???
    (recall FTL has \(|\mathcal{H}|\) mistake bound even if realizable)

• Linear prediction
  • With 01 error, not online learnable (no vanishing regret), even in realizable case
    ➔ Finite VC not enough for online learning, even if realizable
  • Realizable with large \( \ell_2 \) margin: Perceptron guarantee matching margin-based statistical learning
  • Non-realizable, with convex Lipschitz loss and low \( \ell_2 \) norm: FTRL attains regret matching statistical learning
    More generally: convex Lipschitz bounded problems
Convex Lipschitz Problems (w.r.t \( \ell_2 \))

\[
\ell : \mathcal{H} \times \mathcal{Z} \rightarrow \mathbb{R}
\]

- \( \mathcal{H} = \mathbb{R}^d \), or a convex subset of \( \mathbb{R}^d \) (or infinite dimensional)
- \( \mathcal{H} \) if bounded: \( \forall w \in \mathcal{H} \|w\|_2 \leq B \)
- \( \ell(w, z) \) convex and \( G \)-Lipschitz in \( w \): \( |\ell(w, z) - \ell(w', z)| \leq G \|w - w'\|_2 \)
- Supervised learning: \( \ell(w, z) = \text{loss}(\langle w, \phi(x) \rangle, y), G = |\text{loss}'| \cdot \|\phi(x)\|_2 \)

- FTRL: \( w_{t+1} = \arg \min_w \left( \frac{1}{t} \sum_{i=1}^t \ell(w_i, z_i) + \lambda_t \|w\|^2 \right) \)
- For any \( u \in \mathbb{R}^d \):
  \[
  \frac{1}{m} \sum_{t=1}^m (\ell(w_t, z_t) + \lambda_t \|w_t\|_2^2) \leq \frac{1}{m} \sum_{t=1}^m (\ell(u, z_t) + \lambda_t \|u\|_2^2) + \frac{1}{m} \sum_{t=1}^m \frac{2G^2}{t\lambda_t}
  \]
- For any \( \|u\| \leq B \), using \( \lambda_t = \sqrt{2G^2/(B^2 t)} \):
  \[
  \frac{1}{m} \sum_{t=1}^m \ell(w_t, z_t) \leq \frac{1}{m} \sum_{t=1}^m \ell(u, z_t) + \sqrt{\frac{32G^2B^2}{m}}
  \]
Question for Today

• FTRL attains regret $O\left(\sqrt{\frac{G^2 B^2}{m}}\right)$ for convex-Lipschitz-bounded problems.

• “Matches” statistical excess error (“regret” versus best possible expected error)

• But computationally expensive (solve an ERM problem at every iteration) and very non-online-ish (not a simple update of previous iterate)

• Can we attain this regret with a computationally simpler rule?
FTRL for Linear Problems

\[\ell(w, g) = \langle w, g \rangle, \quad g \in G \subset \mathbb{R}^d\]

• FTRL:

\[w_{t+1} = \arg \min_w \frac{1}{t} \sum_{i=1}^{t} \langle w, g_i \rangle + \lambda_t \|w\|^2\]

\[\Rightarrow w_{t+1} = -\frac{1}{2\lambda_t} \sum_{i=1}^{t} g_i = \frac{\lambda_{t-1}(t-1)}{\lambda_t t} w_t - \frac{1}{2\lambda_t} g_t\]

• With \(\lambda_t \propto \frac{1}{t}\), e.g. \(\lambda_t = \frac{\lambda}{t}\):

\[w_{t+1} = w_t - \frac{1}{2\lambda} g_t\]

• In any case: easy to implement incremental rule
  • Only requires storing \(w_t\), not entire history
  • Single vector operation per iteration
FTRL for Linear Problems: Regret

- For $\mathcal{G} = \{g \mid \|g\|_2 \leq G\}$ and $\mathcal{H} = \{w \mid \|w\|_2 \leq B\}$:

- Using $\lambda_t = \frac{\lambda}{t}$ yielding

  $$w_{t+1} = w_t - \frac{1}{2\lambda} g_t$$

  $$\text{Reg}(m) \leq \frac{1}{m} \sum_{t=1}^{m} \left( \frac{\lambda}{t} B^2 + \frac{2G^2}{\lambda} \right) \leq \ln m + 1 \quad \frac{\lambda B^2}{m} + \frac{2G^2}{\lambda} \leq O \left( \sqrt{\frac{B^2 G^2 \log m}{m}} \right)$$

- To avoid log-factor and steps depending on $m$, use $\lambda_t = \sqrt{2G^2/(B^2 t)}$:

  $$w_{t+1} = \sqrt{\frac{t-1}{t}} w_t - \sqrt{\frac{B^2}{8G^2 t}} g_t$$

  $$\text{Reg}(m) \leq \sqrt{\frac{32G^2 B^2}{m}}$$
Linearizing Non-Linear Problems

• For general learning problems, convenient to view as:
  \[ \ell(w, z) = z(w) \]
  where instances \( z \) are functions \( z: \mathcal{H} \rightarrow \mathbb{R} \) (i.e. \( \mathcal{Z} \subset \mathbb{R}^{\mathcal{H}} \))

• Plan:
  • Bound convex \( z(w) \) using linear functions \( \langle g, w \rangle \)
  • Show that low regret on linear functions ensures low regret on \( z(w) \)
  • Conclude: enough to consider FTRL on linear objectives
Sub-Gradients of Convex Functions

• Consider functions over a convex subset \( \mathcal{W} \) of \( \mathbb{R}^d \)

• Definition: \( g \in \mathbb{R}^d \) is a subgradient of a function \( z: \mathcal{W} \to \mathbb{R} \) at \( w_0 \in \mathcal{W} \) iff for all \( w \in \mathcal{W} \), \( z(w) \geq z(w_0) + \langle g, w - w_0 \rangle \)

• The subdifferential \( \partial z(w_0) \) is the set of subgradients at \( w_0 \)

• Claim: A function \( z: \mathcal{W} \to \mathbb{R} \) is convex if and only if it has a subgradient at each point \( w \in \mathcal{W} \)

• Claim: If \( z(w) \) is convex and differentiable at an interior point \( w_0 \in \mathcal{W} \), its unique subgradient at \( w_0 \) is its gradient \( \nabla z(w_0) \)

• At non-differentiable points, there might be multiple sub-gradients

• Claim: A convex function \( z(w) \) is \( G \)-Lipschitz over \( \mathcal{W} \) iff all its subgradients \( g \in \partial z(w) \) at internal points \( w \in \mathcal{W} \) have norm \( \| g \| \leq G \).
Sub-Gradients Are Dual Vectors

• Consider a convex subset $\mathcal{W}$ of vector space $\mathcal{B}$ (e.g. $\mathbb{R}^d$) and functions $z: \mathcal{W} \to \mathbb{R}$
• Recall the dual space $\mathcal{B}^*$ of $\mathcal{B}$ is the vector space of linear functions over $\mathcal{B}$.
• $\phi \in \mathcal{B}^*$ are linear functions $\phi: \mathcal{B} \to \mathbb{R}$, and we denote $\langle \phi, w \rangle = \phi(w)$
• A subgradient of $z: \mathcal{W} \to \mathbb{R}$ at $w_0$ is $g \in \mathcal{B}^*$ s.t.
  $\forall w \in \mathcal{W} z(w) \geq z(w_0) + \langle g, w - w_0 \rangle$
• For $\mathcal{B} = \mathbb{R}^d$, we can think of $\mathcal{B}$ and $\mathcal{B}^*$ and the spaces of row and column vectors.
Dual Norms and Lipschitz

• Recall that for a norm $\|w\|$ over a vector space $\mathcal{B}$, we can define a dual norm over $v \in \mathcal{B}^*$:

$$\|v\|_* = \sup_{\|w\| \leq 1} \langle v, w \rangle = \sup_{\|w\|} \frac{\langle v, w \rangle}{\|w\|}$$

• E.g. over $\mathcal{B} = \mathbb{R}^d$:
  • For $\|w\| = \|w\|_2$, $\|v\|_* = \|v\|_2$
  • For $\|w\| = \|w\|_1$, $\|v\|_* = \|v\|_\infty$
  • For $\|w\| = \|w\|_\infty$, $\|v\|_* = \|v\|_1$
  • For $\|w\| = \|w\|_p$, $\|v\|_* = \|v\|_q$ where $\frac{1}{p} + \frac{1}{q} = 1$

• Claim: for a convex $z: \mathcal{W} \to \mathbb{R}$, $\mathcal{W} \subseteq \mathcal{B}$, $z(w)$ is $G$-Lipschitz over $\mathcal{W}$ w.r.t norm $\|w\|$ iff all subgradients $g \in \partial z(w)$ at internal points $w \in \mathcal{W}$ have norm $\|g\|_*$

Proof: If $\|\nabla z\| \leq G: z(w_1) - z(w_2) \leq z(w_1) - (z(w_1) + \langle \nabla z(w_1), w_2 - w_1 \rangle) \leq \|\nabla z(w_1)\|_*$

$\cdot \|w_2 - w_1\|$

If Lipschitz: $z(w) + \langle \nabla z(w), w + u - w \rangle \leq z(w + u)$,

$\Rightarrow \langle \nabla z(w), u \rangle \leq z(w + u) - z(w) \leq G \|u\|$. Since $w$ is internal, can take $u$ in any direction.
Regret and Linear Lower Bounds

• Consider a general convex online learning problem $\ell(w, z)$ where $w \in \overline{\mathcal{H}} \subseteq \mathcal{B}$ and $z: \overline{\mathcal{H}} \rightarrow \mathbb{R}$

• We will relate regret on this convex problem to regret on the linear online problem $\tilde{\ell}(w, g) = \langle g, w \rangle$

• For any sequence $z_1, \ldots, z_m$ and any online learning rule yielding the predictor sequence $w_1, \ldots, w_m$, consider the sequence of subgradients $g_1 \in \partial z_1(w_1), \ldots, g_t \in \partial z_i(w_i), \ldots, g_m \in \partial z_m(w_m)$

• Claim: for any hypothesis class $\mathcal{H} \subseteq \overline{\mathcal{H}}$:

\[
\left( \frac{1}{m} \sum_{t=1}^{m} \ell(w_t, z_t) - \inf_{u \in \mathcal{H}} \frac{1}{m} \sum_{t=1}^{m} \ell(u, z_t) \right) \leq \left( \frac{1}{m} \sum_{t=1}^{m} \tilde{\ell}(w_t, g_t) - \inf_{u \in \mathcal{H}} \frac{1}{m} \sum_{t=1}^{m} \tilde{\ell}(u, g_t) \right)
\]

Proof: Consider the affine losses $\ell_t(w) = z_t(w_t) + \langle g_t, w - w_t \rangle$. These differ only by a constant (independent of the predictor argument) from the linear losses, and hence have the same regret (difference between online performance and optimal). But $\ell(w_t, z_t) = \tilde{\ell}_t(w_t)$ while $\ell(u, z_t) \geq \tilde{\ell}_t(u)$. 

convex regret \hspace{2cm} linear regret
Reducing Convex to Linear

• Conclusion: we can reduce convex online learning to linear online learning

• Suppose we have a learning rule $A$ that attains regret $Reg_A(m)$ for linear problems $\ell(w, g) = \langle g, w \rangle$ over $g \in \mathcal{G}$ and a hypothesis class $\mathcal{H} \subseteq \mathbb{R}^d$

• Consider the convex problem $\ell(w, z)$ where for all $z \in \mathcal{Z} \subseteq \mathbb{R}^{\overline{\mathcal{H}}}$ and all $w \in \overline{\mathcal{H}}$, $\partial z(w) \subseteq \mathcal{G}$

• We define the learning rule $w_{t+1} = \tilde{A}(z_1, ..., z_t)$
  $= A(\nabla z_1(w_1), \nabla z_2(w_2), ..., \nabla z_t(w_t))$ for any subgradients $\nabla z_i(w_i) \in \partial z_i(w_i)$

  $$Reg_{\tilde{A}}(m) \leq Reg_A(m)$$

• In particular: if we have a learning rule $A$ that attains regret $Reg(m)$ for linear problems over $\mathcal{G} = \{g \mid \|g\|_* \leq G\}$ and hypothesis class $\mathcal{H}$
  $= \{w \mid \|w\| \leq B\}$, then $\tilde{A}$ attains regret $Reg(m)$ for $G$-Lipschitz $B$-Bounded convex problems w.r.t norm $\|w\|$
Online Gradient Descent

• Consider $G$-Lipschitz $B$-bounded convex problems w.r.t. $\|w\|_2$

• We have an easily implementable learning rule for corresponding linear class: FTRL

\[ w_{t+1} = w_t - \frac{1}{2\lambda} g_t \]

• Corresponding rule for convex Lipschitz problems:

\[ w_{t+1} = w_t - \frac{1}{2\lambda} \nabla z_t(w_t) \]

• With above update, $Reg(m) = O \left( \sqrt{\frac{B^2G^2 \log m}{m}} \right)$, or scale $w_t$ to avoid log-factor