Lecture 14:
From Follow the Regularized Leader
To Online Gradient Descent
Question for Today

• FTRL attains regret $O\left(\sqrt{\frac{G^2B^2}{m}}\right)$ 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)

$$w_{t+1} = FTRL(z_1, \ldots, z_t) = \arg\min_w \frac{1}{t} \sum_{i=1}^{t} \ell(w, z_t) + \lambda_t \|w_t\|^2$$

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

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

- **FTRL:**

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

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

- 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

\[ \ell(w, g) = \langle g, w \rangle \]

\( g \in \mathcal{G} = \{g \mid \|g\|_2 \leq G\} \quad \mathcal{H} = \{w \mid \|w\|_2 \leq B\} : \]

\( \Rightarrow \) B-bounded, G-Lipschitz

\[
\text{Reg}(m) \leq \frac{1}{m} \sum_{t=1}^{m} \left( \frac{\lambda_t}{t} B^2 + \frac{2G^2}{\lambda_t} \right) \leq \sqrt{\frac{32G^2B^2}{m}}
\]

\( \lambda_t = \sqrt{\frac{2G^2}{(B^2 t)}} \)

\[
w_{t+1} = \sqrt{\frac{t-1}{t}} w_t - \sqrt{\frac{B^2}{8G^2 t}} g_t
\]
Back to Non-Linear Problems

\[ \ell: \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R} \]

- \( w \mapsto \ell(w, z) \) convex and \( G \)-Lipschitz w.r.t. \( \|w\|_2 \) for every \( z \in \mathcal{Z} \)
- Regret w.r.t. hypothesis class \( \mathcal{W} \subseteq \mathbb{R}^d \) and \( \mathcal{W} \subseteq \{ \|w\|_2 \leq B \} \)

Plan:

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

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

• 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

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

\[
\begin{align*}
z(w) &= \begin{cases} 
-x, & x < 0 \\
x^2, & x \geq 0
\end{cases} \\
\partial z(w) &= \begin{cases} 
{-1}, & w < 0 \\
{2w}, & w > 0
\end{cases}
\]

\[
\partial z(0) = [-1,0]
\]
Sub-Gradients

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

• 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

• 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 (at least one) subgradient at each point $w \in \mathcal{W}$ (i.e. $\partial z(w_0) \neq \emptyset$)
Sub-Gradients

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

- 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

- 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 (at least one) subgradient at each point \( w \in \mathcal{W} \) (i.e. \( \partial z(w_0) \neq \emptyset \) )

- Claim: A convex function \( z: \mathcal{W} \to \mathbb{R} \) is \( G \)-Lipschitz w.r.t. \( \|w\|_2 \) iff all its subgradients \( g \in \partial z(w) \) at internal points \( w \in \mathcal{W} \) have norm \( \|g\|_2 \leq G \).

- Claim: A convex function \( z: \mathcal{W} \to \mathbb{R} \) is \( G \)-Lipschitz w.r.t. \( \|w\| \) iff all its subgradients \( g \in \partial z(w) \) at internal points \( w \in \mathcal{W} \) have norm \( \|g\|_* \leq G \).

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, and so \( \|\nabla z(w)\|_* = \sup_{u} \frac{\langle \nabla z(w), u \rangle}{\|u\|} \leq G \)
Linearizing

• For a general convex $\ell(w, z)$, given $z_1, \ldots, z_m$ and a rule yielding $w_1, \ldots, w_m$ define the linearized problem:

$$
\tilde{\ell}(w, z_i) \overset{\text{def}}{=} \ell(w_i, z_i) + \langle \nabla \ell(w_i, z_i), w - w_i \rangle = \text{const} + \langle g_i, w \rangle
$$

where $g_i = \nabla \ell(w_i, z_i)$ only depends on $z_i, w_i$ independent of $w$.

Given rule $A$ for linear problem, run $A$ on $g_t = \nabla \ell(w_t, z_t)$ i.e. $w_{t+1} = A(g_1, g_2, \ldots, g_t) = A(\nabla \ell(w_1, z_1), \ldots, \nabla \ell(w_t, z_t))$.
Reducing Convex to Linear

convex $\ell(w, z)$

\[ z \in \mathcal{Z} \]

\[ w \in \mathcal{W} \subseteq \mathbb{R}^d \]

\[ \tilde{A}(z_1, ..., z_t) = A(\nabla \ell(w_1, z_1), ..., \nabla \ell(w_t, z_t)) \]

Linear $\ell(w, g) = \langle g, w \rangle$

\[ g \in \mathcal{G} = \{ \nabla \ell(w, z) \mid w \in \mathcal{W}, z \in \mathcal{Z} \} \subset (\mathbb{R}^d)^* \]

\[ w \in \mathcal{W} \subseteq \mathbb{R}^d \]

Learning rule $A(g_1, ..., g_m)$

\[
\text{Reg}_{\tilde{A}}(m) \leq \text{Reg}_A(m)
\]

In particular: if $A$ that attains regret $\text{Reg}(m)$ for linear problems over

\[ \mathcal{G} = \{ g \mid \|g\|_* \leq G \} \]

and hypothesis class \[ \mathcal{W} = \{ w \mid \|w\| \leq B \} \], then $\tilde{A}$ attains $\text{Reg}(m)$ for $G$-Lipschitz $B$-Bounded convex problems w.r.t $\|w\|$

$\Rightarrow$ FTRL on $\nabla \ell(w_t, z_t)$ attains $\text{Reg}_{\text{FTRL}}(m) \leq \sqrt{\frac{32B^2G^2}{m}}$ on

$G$-Lipschitz $B$-Bounded convex problems w.r.t. $\|w\|_2$
Follow the Regularized Linearized Leader
aka Online Gradient Descent

- $\ell(w, z)$ convex and $G$-Lipschitz w.r.t. $\|w\|_2$ for every $z \in Z$
- $\mathcal{W} \subseteq \{ \|w\|_2 \leq B \}$

Follow the Regularized Linearized Leader:

$$w_{t+1} \leftarrow \arg \min_w \frac{1}{t} \sum_{i=1}^t \langle \nabla \ell(w_i, z_i), w \rangle + \lambda_t \|w\|_2^2$$

$$= \frac{\lambda_{t-1}(t-1)}{\lambda_t t} w_t - \frac{1}{2\lambda_t t} \nabla \ell(w_t, z_t) = \sqrt{\frac{t-1}{t}} w_t - \sqrt{\frac{B^2}{8G^2 t}} \nabla \ell(w_t, z_t)$$

Using $\lambda_t = \frac{\lambda}{t}$:

$$w_{t+1} \leftarrow w_t - \frac{1}{2\lambda} \nabla \ell(w_t, z_t)$$

Using stability analysis, $O \left( \sqrt{\frac{B^2 G^2 \log m}{m}} \right)$ regret. Actually, no log factor.
Answer 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 and not a simple update rule
• Can we attain this regret with a computationally simpler rule?
• FTRLL/OGD attains same regret using simple and cheap update rule
“The” Convex Lipschitz Bounded Problem

• We can view instances as functions \( z: \overline{\mathcal{H}} \rightarrow \mathbb{R} \) specifying the loss and:
  \[ \ell(w, z) = z(w) \]
i.e., at each round predict \( w_t \), receive function \( z_t \) and pay \( z_t(w_t) \)

• Learning problem then specified by:
  • \( \mathcal{Z} \subset \overline{\mathcal{H}}^{\mathbb{R}} \) — the class of functions to be learning/optimized
  • \( \mathcal{H} \subset \overline{\mathcal{H}} \) — the hypothesis class we compare performance to

• Learning problem of all convex Lipschitz bounded functions:
  • \( \overline{\mathcal{H}} = \mathbb{R}^d \)
  • \( \mathcal{Z} = \{ z: \mathbb{R}^d \rightarrow \mathbb{R} \text{ convex and } G-\text{Lipschitz} \} \)
  • \( \mathcal{H} = \mathcal{W} = \{ w \mid \|w\| \leq B \} \)

• What we showed: FLRLL/OGD learns the convex Lipschitz bounded problem.
• Recall Perceptron update:
  \[ w_{t+1} \leftarrow w_t + \left[ [y_t \langle w, x_t \rangle \leq 0] \right] \cdot y_t x_t \]

• Can be viewed as OGD on \( \ell(w, (x, y)) = hinge_0(y\langle w, x \rangle) \)
  - \( L_{S_{\text{mrg}}} (w) = 0 \Rightarrow L_{S_{\text{hinge}0}} (w) = 0 \)
  - But: doesn’t upper bound 01-loss!

• Instead:
  - Ignore correctly classified points
  - View as OGD on \( \ell(w, (x, y)) = hinge(y\langle w, x \rangle) \)

• Claim: if \( A \) achieves mistake bound \( M \), and we run \( A \) only on mistakes,
  \[ h_{t+1} = \tilde{A}(z_1, \ldots, z_t) = A(\{z_i\}_{t=1}^{i}, h_i(x_i) \neq y_i) \]
  then \( \tilde{A} \) makes at most \( M \) mistakes