Computational and Statistical Learning Theory

TTIC 31120

Prof. Nati Srebro

Lecture 16:

From Follow the Regularized Leader to Online Gradient Descent and the Perceptron Rule

Question for Today

FTRL has regret $O\left(\sqrt{\frac{G^2B^2}{m}}\right)$ for cvnx Lipschitz bounded problems wrt $\|w\|_2$

("Matches" statistical excess error)

But computationally expensive very non-online-ish
(not a simple update of previous iterate)

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

- Solve an ERM-type problem at every iteration
- Need to store all previous examples $(z_1, \dots, z_t)$, i.e. $O(md)$ memory (vs $O(d)$ for Perceptron)

Can we attain this regret with a computationally simpler rule?

FTRL for Linear Problems

$$\ell(w, g) = \langle g, w \rangle, \qquad g \in \left(\mathbb{R}^d\right)^*$$

market behavior
$g[i]$= $-$(return on stock i)

investment portfolio
$w[i]$=holding in stock i

FTRL for Linear Problems

$$\ell(w, g) = \langle g, w \rangle, \qquad g \in \left(\mathbb{R}^d\right)^*$$

- FTRL:

$$w_{t+1} = \arg\min_w \tfrac{1}{t} \textstyle\sum_{i=1}^{t} \langle g_i, w \rangle + \lambda_t \|w\|^2$$

$$= \arg\min_w \langle \frac{1}{t} \sum g_i, w \rangle + \lambda_t \|w\|^2$$

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

- 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

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

$g \in \mathcal{G} = \{g \mid \|g\|_2 \leq G\}$

G-Lipschitz

$\mathcal{H} = \{w \mid \|w\|_2 \leq B\}$:

B-bounded

$$\Rightarrow w_{t+1} = \frac{\lambda_{t-1}(t-1)}{\lambda_t t} w_t - \frac{1}{2\lambda_t t} g_t = \sqrt{\frac{t-1}{t}} w_t - \sqrt{\frac{B^2}{8G^2 t}} g_t$$

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

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

Back to Non-Linear Problems

$$\ell: \mathbb{R}^d \times \mathcal{Z} \rightarrow \mathbb{R}$$

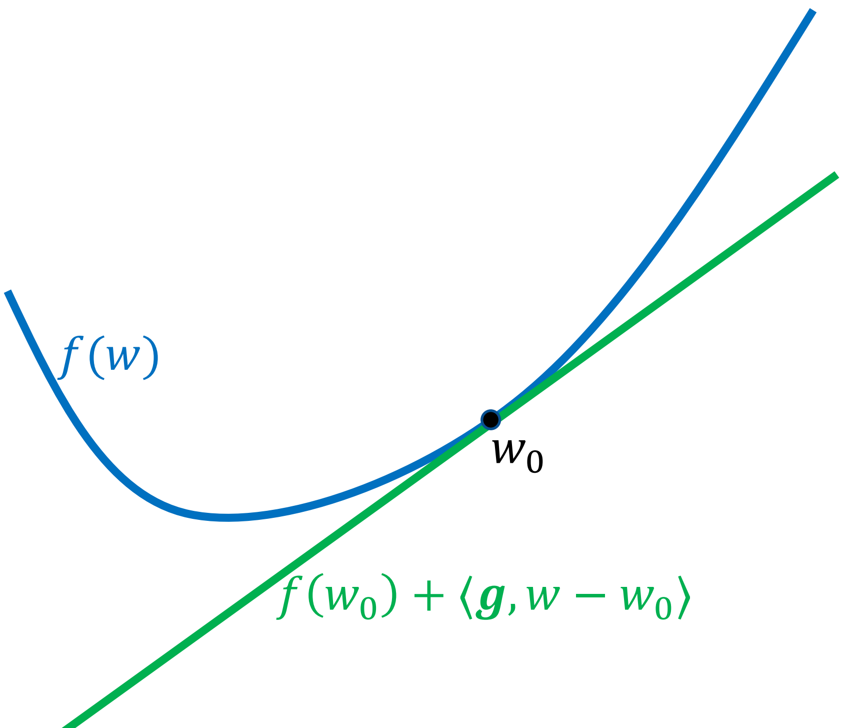
- $w \mapsto \ell(w, z)$ convex and G-Lipschitz w.r.t. $\|w\|_2$ for every $z \in \mathcal{Z}$
- Regret w.r.t. hypothesis class $\mathcal{W} \subseteq \mathbb{R}^d$ and $\mathcal{W} \subseteq \{\ \|w\|_2 \leq B\ \}$

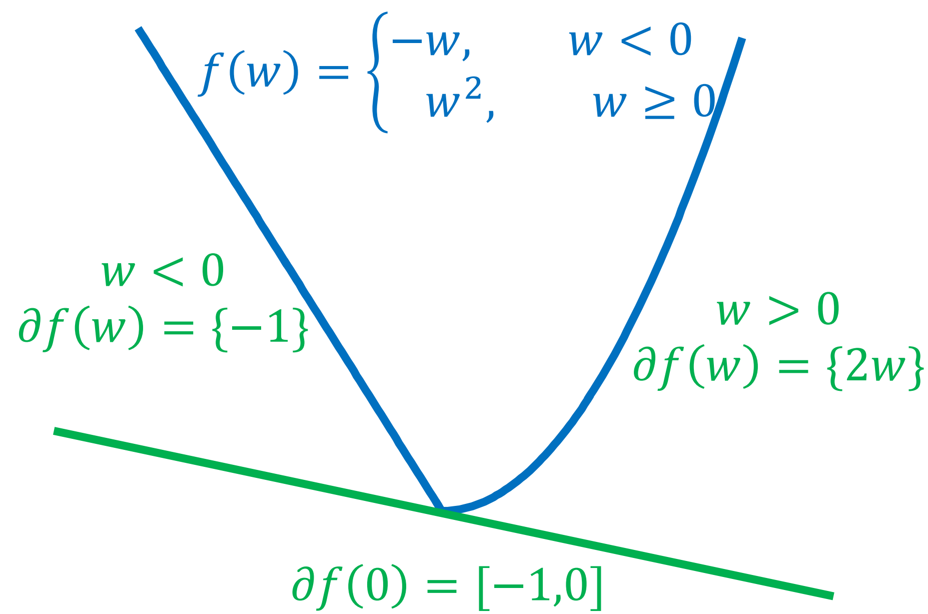
- Plan:
  - Bound convex $\ell(w, z)$ using linear functions $\langle g, w \rangle$
  - Show low regret on linear functions ensures low regret on $\ell(w, z)$
  - Conclude: enough to consider FTRL on linear objectives

Sub-Gradients

e.g. $\mathcal{W} = \mathbb{R}^d$
$\mathcal{W}^* \cong \mathbb{R}^d$

- Definition: $g \in \mathcal{W}^*$ is a subgradient of a function $f : \mathcal{W} \to \mathbb{R}$ at $w_0 \in \mathcal{W} \subseteq \mathbb{R}^d$ iff for all $w \in \mathcal{W}$, $f(w) \geq f(w_0) + \langle \boldsymbol{g}, w - w_0 \rangle$
- Claim: If $f(w)$ is convex and differentiable at an interior point $w_0 \in \mathcal{W}$, its unique subgradient at w_0 is its gradient $\nabla f(w_0)$
- At non-differentiable points, there might be multiple sub-gradients
- The subdifferential $\partial f(w_0)$ is the set of subgradients at w_0

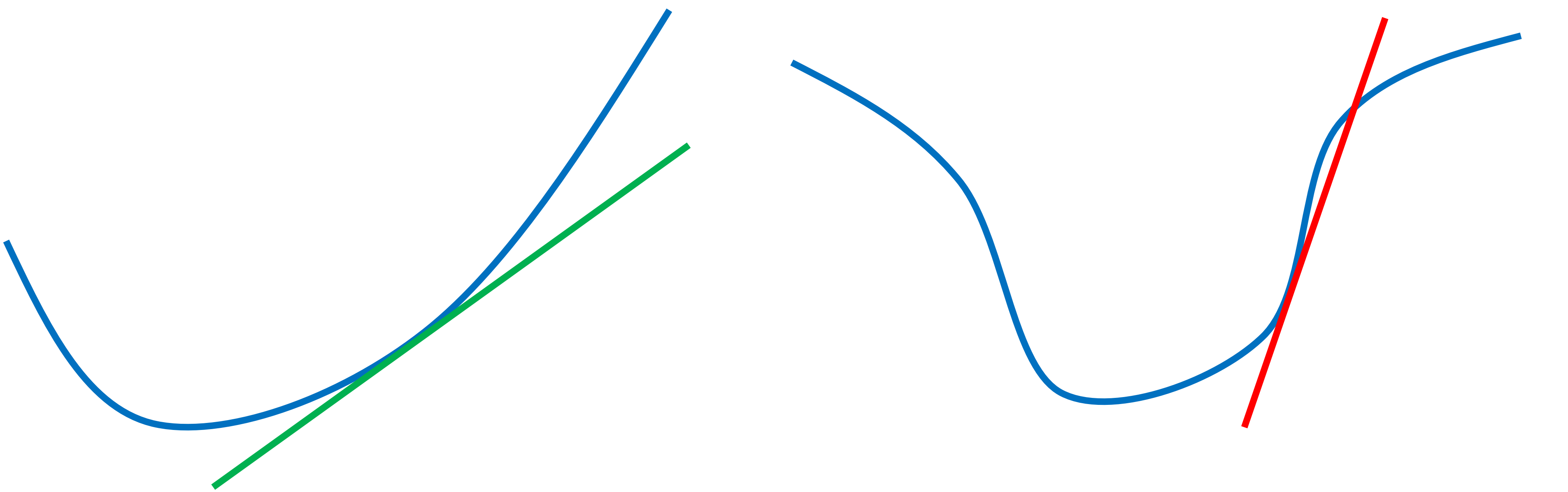




Sub-Gradients

e.g. $\mathcal{W} = \mathbb{R}^d$
$\mathcal{W}^* \cong \mathbb{R}^d$

- Definition: $g \in \mathcal{W}^*$ is a subgradient of a function $f: \mathcal{W} \to \mathbb{R}$ at $w_0 \in \mathcal{W} \subseteq \mathbb{R}^d$ iff for all $w \in \mathcal{W}$, $f(w) \geq f(w_0) + \langle \boldsymbol{g}, w - w_0 \rangle$
- Claim: If $f(w)$ is convex and differentiable at an interior point $w_0 \in \mathcal{W}$, its unique subgradient at w_0 is its gradient $\nabla f(w_0)$
- At non-differentiable points, there might be multiple sub-gradients
- The subdifferential $\partial f(w_0)$ is the set of subgradients at w_0
- Claim: A function $f: \mathcal{W} \to \mathbb{R}$ is convex if and only if it has (at least one) subgradient at each point $w \in \mathcal{W}$ (i.e. $\partial f(w_0) \neq \emptyset$)



Sub-Gradients

e.g. $\mathcal{W} = \mathbb{R}^d$
$\mathcal{W}^* \cong \mathbb{R}^d$

- Definition: $g \in \mathcal{W}^*$ is a subgradient of a function $f: \mathcal{W} \to \mathbb{R}$ at $w_0 \in \mathcal{W} \subseteq \mathbb{R}^d$ iff for all $w \in \mathcal{W}$, $f(w) \geq f(w_0) + \langle \boldsymbol{g}, w - w_0 \rangle$
- Claim: If $f(w)$ is convex and differentiable at an interior point $w_0 \in \mathcal{W}$, its unique subgradient at w_0 is its gradient $\nabla f(w_0)$
- At non-differentiable points, there might be multiple sub-gradients
- The subdifferential $\partial f(w_0)$ is the set of subgradients at w_0
- Claim: A function $f: \mathcal{W} \to \mathbb{R}$ is convex if and only if it has (at least one) subgradient at each point $w \in \mathcal{W}$ (i.e. $\partial f(w_0) \neq \emptyset$)
- Claim: A convex function $f: \mathcal{W} \to \mathbb{R}$ 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 $f: \mathcal{W} \to \mathbb{R}$ 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 $\|\nabla f\| \leq G$: $f(w_1) - f(w_2) \leq f(w_1) - (f(w_1) + \langle \nabla f(w_1), w_2 - w_1 \rangle \leq \|\nabla f(w_1)\|_* \cdot \|w_2 - w_1\|$

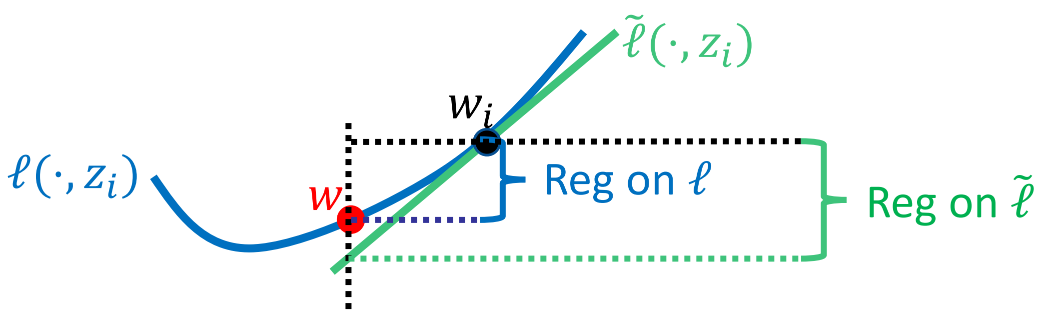
If Lipschitz: $f(w) + \langle \nabla f(w), w + u - w \rangle \leq f(w + u)$ ➔ $\langle \nabla f(w), u \rangle \leq f(w+u) - f(w) \leq G\|u\|$.
Since w is internal, can take u in any direction, and so $\|\nabla f(w)\|_* = \sup_u \frac{\langle \nabla f(w), u \rangle}{\|u\|} \leq G$

Linearizing

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

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

only depends on z_i, w_i
independent of w



$$\left(\frac{1}{m}\sum_{t=1}^{m} \ell(w_t, z_t) - \inf_{w \in \mathcal{W}} \frac{1}{m}\sum_{t=1}^{m} \ell(w, z_t)\right) \le \left(\frac{1}{m}\sum_{t=1}^{m} \tilde{\ell}(w_t, z_t) - \inf_{w \in \mathcal{W}} \frac{1}{m}\sum_{t=1}^{m} \tilde{\ell}(w, z_t)\right)$$

$$= \left(\frac{1}{m}\sum_{t=1}^{m} \langle g_t, w_t \rangle - \inf_{w \in \mathcal{W}} \frac{1}{m}\sum_{t=1}^{m} \langle g_t, w \rangle\right) \le Reg_A(m)$$

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

Reducing Convex to Linear

convex $\ell(w, z)$
$\|w\| \leq B$
G-Lipschitz wrt $\|w\|$

linear $\bar{\ell}(w, g) = \langle g, w \rangle$
$\|w\| \leq B$ (same hypothesis class of w)
$\|g\|_* = \|\nabla \ell(w, z)\|_* \leq G$

$\tilde{\boldsymbol{A}}(z_1, \ldots, z_t) = \boldsymbol{A}\left(\nabla \ell(w_1, z_1), \ldots, \nabla \ell(w_t, z_t)\right)$

Learning rule $\boldsymbol{A}(g_1, \ldots, g_m)$

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

In particular: if A that attains regret $Reg(m)$ for linear problems over $\{g \mid \|g\|_* \leq G\}$ and hypothesis class $\{ w \mid \|w\| \leq B \}$, then $\tilde{A}$ attains $Reg(m)$ for G-Lipschitz B-Bounded convex problems w.r.t $\|w\|$

➔ FTRL on $\nabla \ell(w_t, z_t)$ attains $Reg_{\widetilde{FTRL}}(m) \leq \sqrt{\frac{32B^2G^2}{m}}$ 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 $z \in \mathcal{Z}$
- $\mathcal{W} \subseteq \{\ \|w\|_2 \leq B\ \}$

Follow the Regularized Linearized Leader:

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

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

$\lambda_t = \sqrt{2G^2/(B^2 t)}$ to achieve
$Reg(m) \leq \sqrt{\frac{32B^2G^2}{m}}$

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, $O\left(\sqrt{\frac{B^2G^2\log m}{m}}\right)$ regret. Actually, no log factor.

Answer for Today

- FTRL attains regret $O\left(\sqrt{\frac{G^2B^2}{m}}\right)$ for convex-Lipschitz-bounded problems.
- But not a simple update
  - Need $O(md)$ memory to keep track of all previous examples
  - Need to solve ERM-like problem at each step

- Can we attain this regret with a computationally simpler rule?

- **FTRLL/OGD attains same regret using simple and cheap update rule**

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

- What about convex-Lipschitz-bounded w.r.t. other $\|w\|$?
- But first…

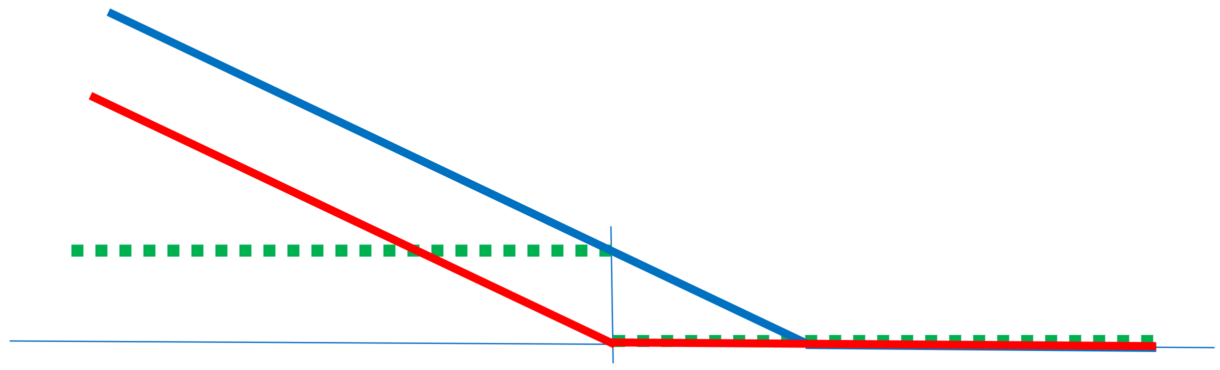
…back to the Perceptron

- Recall Perceptron update:
$$w_{t+1} \leftarrow w_t + [[y_t\langle w, x_t\rangle \leq 0]] \cdot y_t x_t$$

- Can be viewed as OGD on $\ell\big(w,(x,y)\big) = hinge_0(y\langle w,x\rangle)$
  - $L_S^{mrg}(w) = 0 \rightarrow L_S^{hinge_0}(w) = 0$
  - But: doesn't upper bound 01-loss!

$hinge_0(y\langle w,x\rangle) = [-y\langle w,x\rangle]_+$

- Can get same guarantee with OGD on $\ell\big(w,(x,y)\big) = [1 - y\langle w,x\rangle]_+$
  - "Aggressive Perceptron": $w_{t+1} \leftarrow w_t + [[y_t\langle w, x_t\rangle \leq \mathbf{1}]] \cdot y_t x_t$



…back to the Perceptron

- Recall Perceptron update:

$$w_{t+1} \leftarrow w_t + [[y_t\langle w, x_t\rangle \leq 0]] \cdot y_t x_t$$

- Can be viewed as OGD on $\ell\big(w,(x,y)\big) = hinge_0(y\langle w,x\rangle)$
  - $L_S^{mrg}(w) = 0 \Rightarrow L_S^{hinge_0}(w) = 0$
  - But: doesn't upper bound 01-loss!

$hinge_0(y\langle w,x\rangle) = [-y\langle w,x\rangle]_+$

- Can get same guarantee with OGD on $\ell\big(w,(x,y)\big) = [1 - y\langle w,x\rangle]_+$
  - "Aggressive Perceptron": $w_{t+1} \leftarrow w_t + \big[[y_t\langle w, x_t\rangle \leq \mathbf{1}]\big] \cdot y_t x_t$

- Instead:
  - Ignore correctly classified points
  - View as OGD on $\ell\big(w,(x,y)\big) = hinge(y\langle w,x\rangle) = [1 - y\langle w,x\rangle]_+$

- Claim: if A achieves mistake bound M, and we run A only on mistakes,

$$h_{t+1} = \tilde{A}(z_1,\dots,z_t) = A\big(\{z_i\}_{t=1..i,h_i(x_i)\neq y_i}\big)$$

then $\tilde{A}$ makes at most M mistakes

Convex Lipschitz Problems

$$\ell : \overline{\mathcal{H}} \times \mathcal{Z} \to \mathbb{R}$$

- $\overline{\mathcal{H}} \subseteq \mathcal{B}$ convex subset of normed vector space, e.g. $\mathcal{B} = \mathbb{R}^d$
- $\mathcal{H} \subseteq \overline{\mathcal{H}}$ is bounded: $\forall_{w \in \mathcal{H}} \|w\| \leq B$
- $\ell(w, z)$ convex and G-Lipschitz w.r.t $\|w\|$:
  $\forall_{z \in \mathcal{Z}, w, w' \in \overline{\mathcal{H}}} |\ell(w, z) - \ell(w', z)| \leq G\|w - w'\|$ or $\|\nabla \ell(w, z)\|_* \leq G$

  E.g. supervised learning: $\ell(w, z) = loss(\langle w, \phi(x) \rangle, y)$,
  $\|\nabla \ell(w, z)\|_* = \|loss'(..) \cdot \phi(x)\|_* = |loss'| \cdot \|\phi(x)\|_* \leq G$

- Need $\Psi(w) \geq 0$ which is α-strongly convex w.r.t. $\|w\|$ on $\overline{\mathcal{H}}$

$$FTRL(z_1, \dots, z_t) = \arg\min_{w \in \overline{\mathcal{H}}} \frac{1}{t} \sum_{i=1}^{t} \ell(w, z_i) + \lambda_t \Psi(w)$$

using $\lambda_t = \sqrt{\frac{2G^2}{\alpha B^2 t}}$, $\boldsymbol{Reg}(\mathbf{FTRL}) \leq \sqrt{\frac{\mathbf{32}\boldsymbol{G}^{\mathbf{2}}\tilde{\boldsymbol{B}}^{\mathbf{2}}}{\boldsymbol{\alpha m}}}$ where $\sup_{w \in \mathcal{H}} \Psi(w) \leq \tilde{B}^2$

Linearized FTRL

$$w_{t+1} = \arg\min_{w \in \mathcal{H}} \frac{1}{t} \sum_{i=1}^{t} \langle \nabla \ell(w_i, z_i), w \rangle + \lambda_t \Psi(w)$$

$$= \arg\min_{w \in \mathcal{H}} \left\langle \underbrace{\frac{1}{t} \sum_{i=1}^{t} \nabla \ell(w_i, z_i)}_{v_t}, w \right\rangle + \lambda_t \Psi(w)$$

- Same regret as FTRL!

$$\sqrt{\frac{32 G^2 \tilde{B}^2}{\alpha m}}$$

- Only need to keep tack of sum of gradient $v_t = \sum_{i=1}^{t} \nabla \ell(w_i, z_i)$

$$v_t = v_{t-1} + \nabla \ell(w_t, z_t)$$

$$w_{t+1} = \arg\min_{w \in \mathcal{H}} \langle v_t, w \rangle + \lambda_t \Psi(w)$$

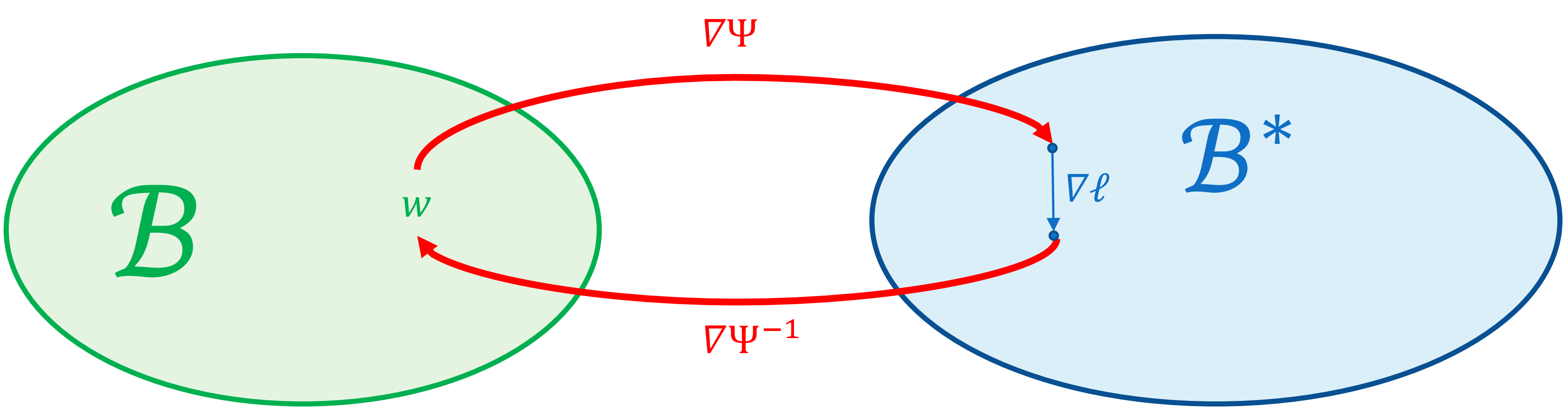
Linearized FTRL

$$w_{t+1} = \arg\min_{w \in \overline{\mathcal{H}}} \frac{1}{t} \sum_{i=1}^{t} \langle \nabla \ell(w_i, z_i), w \rangle + \lambda_t \Psi(w)$$

- If $\overline{\mathcal{H}} = \mathcal{B}$: $\quad 0 = \frac{1}{t} \sum_{i=1}^{t} \nabla \ell(w_i, z_i) + \lambda_t \nabla \Psi(w_{t+1})$

$$\Rightarrow w_{t+1} = \nabla \Psi^{-1}\left(-\frac{1}{\lambda_t t} \sum_{i=1}^{t} \nabla \ell(w_i, z_i)\right)$$

$$\Rightarrow w_{t+1} = \nabla \Psi^{-1}\left(\frac{\lambda_{t-1}(t-1)}{\lambda_t t} \nabla \Psi(w_t) - \frac{1}{\lambda_t t} \nabla \ell(w_t, z_t)\right)$$



Linearized FTRL (aka "Dual Averaging")

$$w_{t+1} = \arg\min_{w \in \overline{\mathcal{H}}} \frac{1}{t} \sum_{i=1}^{t} \langle \nabla \ell(w_i, z_i), w \rangle + \lambda_t \Psi(w)$$

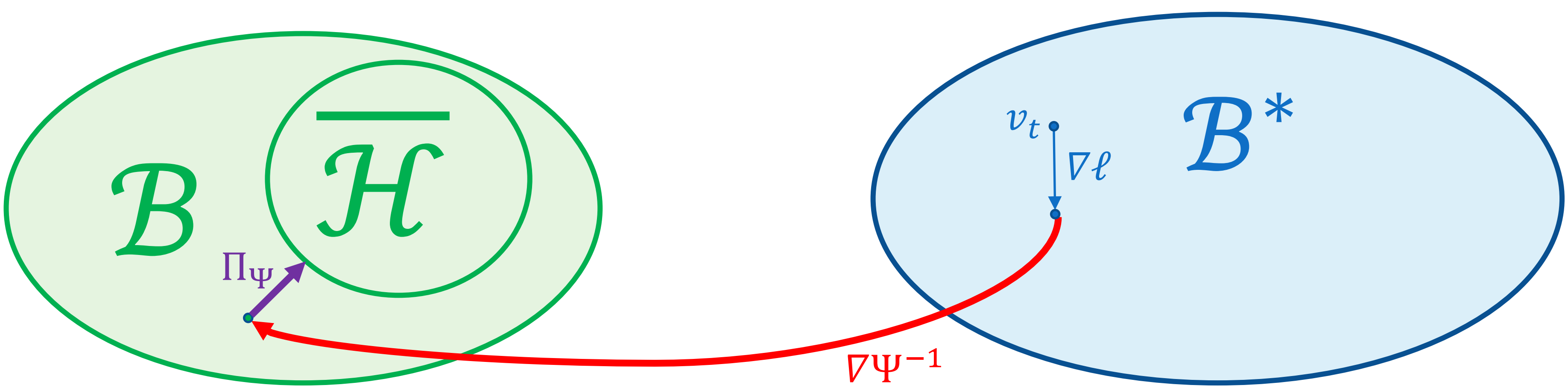
- If $\overline{\mathcal{H}} = \mathcal{B}$: $w_{t+1} = \nabla \Psi^{-1}\left(\frac{\lambda_{t-1}(t-1)}{\lambda_t t} \nabla \Psi(w_t) - \frac{1}{\lambda_t t} \nabla \ell(w_t, z_t) \right)$
- If $\overline{\mathcal{H}} \subset \mathcal{B}$: $w_{t+1} = \Pi_{\Psi}^{\overline{\mathcal{H}}}\left(\nabla \Psi^{-1}\left(-\frac{1}{\lambda_t} v_t \right) \right)$ $v_t = v_{t-1} + \nabla \ell(w_t, z_t)$

Where: $\Pi_{\Psi}^{\mathcal{W}}(w) = \arg\min_{w' \in \mathcal{W}} D_{\Psi}(w' || w)$

Beyond required scope of course

Bergman Divergence: $D_{\Psi}(w' || w) = \Psi(w') - (\Psi(w) + \langle \nabla \Psi(w), w' - w \rangle)$

Proof: $\Pi\left(\nabla \Psi^{-1}(v)\right) = \arg\min_{w' \in \overline{\mathcal{H}}} \Psi(w') - \langle \nabla \Psi\left(\nabla \Psi^{-1}\left(\frac{-v}{\lambda_t}\right)\right), w' \rangle = \arg\min_{w \in \overline{\mathcal{H}}} \langle v, w \rangle + \lambda_t \Psi(w)$



How to Choose Ψ

Instantaneous loss $\ell(w, z) = loss(\langle w, \phi(x)\rangle, y)$

- G-Lipschitz w.r.t. $\|w\|$, i.e. $\|\nabla\ell\| \leq |loss'| \cdot \|\phi(x)\|_* \leq G$
- Compete with $w \in \mathcal{H}$

- Find Ψ which is:
  - α-strongly convex w.r.t. $\|w\|$ on $\bar{\mathcal{H}}$
  - $\forall_{w\in\mathcal{H}} 0 \leq \Psi(w) \leq \tilde{B}^2$
  - Easy to compute $\nabla\Psi$, $\nabla\Psi^{-1}$ and if needed also $\Pi_{\Psi}^{\bar{\mathcal{H}}}$

- Regret: $O\left(\sqrt{\frac{\tilde{B}^2 G^2}{\alpha m}}\right)$

Example: $\|w\|_2$

- $\Psi(w) = \frac{1}{2}\|w\|_2^2$ is 1-strongly convex wrt $\|w\|_2$
- Regret: $O\left(\sqrt{\frac{\|w\|_2^2 \|\nabla\|_2^2}{m}}\right)$
- $\nabla\Psi(w) = w^\top$, $\nabla\Psi^{-1}(v) = v^\top$
- FTRLL/OGD: $w_{t+1} = \frac{\lambda_{t-1}(t-1)}{\lambda_t t} w_t - \frac{1}{\lambda_t t} \nabla\ell(w_t, z_t)$

Example: $\|w\|_Q = \sqrt{w^T Q w}$

- $\Psi(w) = \frac{1}{2} w^T Q w$ is 1-strongly convex w.r.t $\|w\|_Q$
- $\|v\|_* = \|v\|_{Q^{-1}} = \sqrt{v^T Q^{-1} v}$
- Regret: $O\left(\sqrt{\frac{(w^T Q w)\left((\nabla \ell)^\top Q^{-1} (\nabla \ell)\right)}{m}}\right)$

- $\nabla \Psi(w) = Qw,\ \nabla \Psi^{-1}(v) = Q^{-1} v$

- Pre-conditioned OGD: $w_{t+1} = \frac{(t-1)\lambda_{t-1}}{t\lambda_t} w_t - Q^{-1} \nabla \ell(w_t, z_t)$

Example: $\|w\|_p$

- $\Psi(w) = \frac{1}{2}\|w\|_p^2$ is $(p-1)$-strongly convex w.r.t. $\|w\|_p$
- Regret: $O\left(\sqrt{\frac{\|w\|_p^2 \|\nabla \ell\|_q^2}{(p-1)m}}\right)$

- $\nabla\Psi(w)[i] = \|w\|_p^{2-p} |w[i]|^{p-1} sign(w[i])$
- $\nabla\Psi^{-1}(v)[i] = \frac{|v[i]|^{q-1} sign(v[i])}{\|v\|_q^{q-2}}$, where $\frac{1}{p} + \frac{1}{q} = 1$

- Explodes as $p \to 1$, what about $\|w\|_1$?

$$\mathcal{H} = \overline{\mathcal{H}} = \{\, w \in \mathbb{R}^{\boldsymbol{d}} \mid w \geq 0, \|\boldsymbol{w}\|_{\boldsymbol{1}} = 1 \,\}$$

- $\Psi(w) = \sum_i w[i] \log \frac{w[i]}{1/d} = \log d + \sum_i w[i] \log w[i]$
- For $w \in \mathcal{H}$: $0 \leq \Psi(w) \leq \log d$
- Claim: $\Psi(w)$ is 1-strongly convex w.r.t. $\|w\|_1$ on $\overline{\mathcal{H}}$
- Regret: $O\left(\sqrt{\frac{\|\nabla \ell\|_\infty^2 \log d}{m}}\right)$

- $\nabla\Psi(w)[i] = (\log w[i]) + 1$
- $\nabla\Psi^{-1}(v)[i] = e^{v[i]-1}$
- $w_{t+1} = \arg\min_{w\in\mathcal{H}} \langle v_t, w\rangle + \lambda_t \Psi(w) = \Pi_{\Psi}^{\sum w[i]=1}\left(\nabla\Psi^{-1}(v_t)\right)$

$$\Rightarrow w_{t+1}[i] = \frac{e^{-\frac{1}{t\lambda_t} v_t[i]}}{\sum_i e^{-\frac{1}{t\lambda_t} v_t[i]}}$$

$$v_t = \sum_{i=1}^{t} \nabla\ell(w_i, z_i)$$

$$\mathcal{H} = \overline{\mathcal{H}} = \{ w \in \mathbb{R}^{\boldsymbol{d}} \mid w \geq 0, \|\boldsymbol{w}\|_1 = 1 \}$$

$v_t = \sum_{i=1}^{t} \nabla \ell(w_i, z_i)$

$$w_{t+1}[i] \propto e^{-\frac{1}{t\lambda_t} v_t[i]} = e^{-\frac{1}{(t-1)\lambda_{t-1}} v_{t-1}[i] - \frac{1}{\lambda} \nabla \ell(w_t, z_t)[i]} \propto w_t[i] e^{-\frac{1}{\lambda} \nabla \ell(w_t, z_t)[i]}$$

with $\lambda_t = \frac{\lambda}{t}$, and so $\frac{1}{t\lambda_t} v_t = \frac{1}{(t-1)\lambda_{t-1}} v_{t-1} + \frac{1}{\lambda} \nabla \ell(w_i, z_i)$

Normalized Exponentiated Gradient (EG)

- $w_1[i] = \frac{1}{d}$
- $w_{t+1}[i] = \frac{w_t[i] e^{-\frac{1}{\lambda} \nabla \ell(w_t, z_t)[i]}}{\sum_j w_t[j] e^{-\frac{1}{\lambda} \nabla \ell(w_t, z_t)[j]}}$

Regret: $O\left(\sqrt{\frac{\|\nabla \ell\|_\infty^2 \log d}{m}}\right)$

Our stability-based analysis gives $O\left(\sqrt{\|\nabla \ell\|_\infty^2 \log d \log m / m}\right)$. We can avoid log-factor with $\lambda_t = \lambda/\sqrt{t}$, but then updates less nice. Alt analysis avoids log factor for EG as above.

Only realizable
(all others also agnostic)

	Online	Statistical
Finite Cardinality	$\log \lvert\mathcal{H}\rvert$ Halving	$\log \lvert\mathcal{H}\rvert$ ERM
Finite Dimension	∞	$VCdim$ ERM
Scale Sensitive Convex	$\lVert w\rVert_2^2 \lVert\nabla\ell\rVert_2^2$ (L)FTRL / OGD $\lVert \mathrm{w}\rVert_1^2 \lVert\nabla\ell\rVert_\infty^2 \log(d)$ (L)FTRL / EG $\Psi(w) \lVert\nabla\ell\rVert_*^2$ (L)FTRL	$\lVert w\rVert_2^2 \lVert\nabla\ell\rVert_2^2$ RERM $\lVert \mathrm{w}\rVert_1^2 \lVert\nabla\ell\rVert_\infty^2 \log(d)$ Boosting / RERM $\Psi(w) \lVert\nabla\ell\rVert_*^2$ RERM

Back to Finite Cardinality

- Consider a finite cardinality hypothesis class $\mathcal{H}$ and bounded loss $0 \leq loss \leq 1$ (e.g. 0/1 error)
- HALVING: regret $\frac{\log|\mathcal{H}|}{m}$ wr.t. 0/1 error ($\frac{\#mistakes}{m}$) in the realizable case
- What about agnostic case? Or general bounded loss?
- Solution: convexification
- Linear loss over $\mathbb{R}^{\mathcal{H}}$, with each coordinate corresponding to a $h \in \mathcal{H}$

$$\ell(w,(x,y)) = \langle w, g(x,y)\rangle \quad \text{with} \quad g(x,y)[h] = loss(h(x);y)$$

- For $e_h = (0, \ldots 0,1,0, \ldots, 0)$, $\quad \ell\big(e_h,(x,y)\big) = loss(h(x);y)$
- Hypothesis class becomes: $\{e_h | h \in \mathcal{H}\}$, non-convex!
- Improper learning with

$$\{e_h | h \in \mathcal{H}\} \subseteq \overline{\mathcal{H}} = \left\{ w \in \mathbb{R}^d \mid w \geq 0, \|w\|_1 = 1 \right\}$$

- $\|g(x,y)\|_\infty \leq \sup loss \leq 1$
- Use normalized EG algorithm

Multiplicative Weights Algorithm

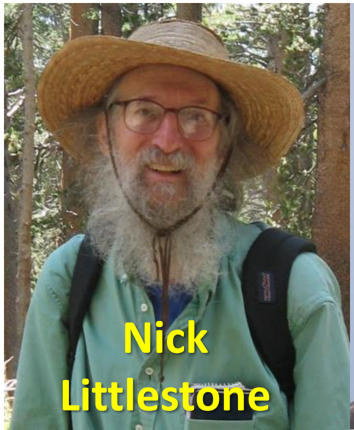
- $w_1[h] = \frac{1}{|\mathcal{H}|}$

At round t:

- Receive x_t
- Pick hypothesis h w.p. $w_t[h]$,
  use it to predict $\hat{y}_t = h(x_t)$, suffer loss $loss(h(x_t), y_t) = g_t[h]$
  ➔ expected loss = $\mathbb{E}_{h \sim w_t} g_t[h] = \langle w_t, g_t \rangle$
- Receive y_t
- $w_{t+1}[h] = \frac{w_t[h] e^{-\frac{1}{\lambda} g_t[h]}}{\sum_j w_t[j] e^{-\frac{1}{\lambda} g_t[j]}}$

Loss if using hypothesis h on (x_t, y_t)

- The expected regret of MW is $O\left(\sqrt{\frac{\log|\mathcal{H}|}{m}}\right)$



Also agnostic!

	Online	Statistical
Finite Cardinality	$\log\|\mathcal{H}\|$ Halving	$\log\|\mathcal{H}\|$ ERM
Finite Dimension	∞	$VCdim$ ERM
Scale Sensitive Convex	$\|\|w\|\|_2^2\|\|\nabla\ell\|\|_2^2$ (L)FTRL / OGD $\|\|\mathrm{w}\|\|_1^2\|\|\nabla\ell\|\|_\infty^2\log(d)$ (L)FTRL / EG $\Psi(w)\|\|\nabla\ell\|\|_*^2$ (L)FTRL	$\|\|w\|\|_2^2\|\|\nabla\ell\|\|_2^2$ RERM $\|\|\mathrm{w}\|\|_1^2\|\|\nabla\ell\|\|_\infty^2\log(d)$ Boosting / RERM $\Psi(w)\|\|\nabla\ell\|\|_*^2$ RERM