Computational and Statistical Learning Theory

TTIC 31120

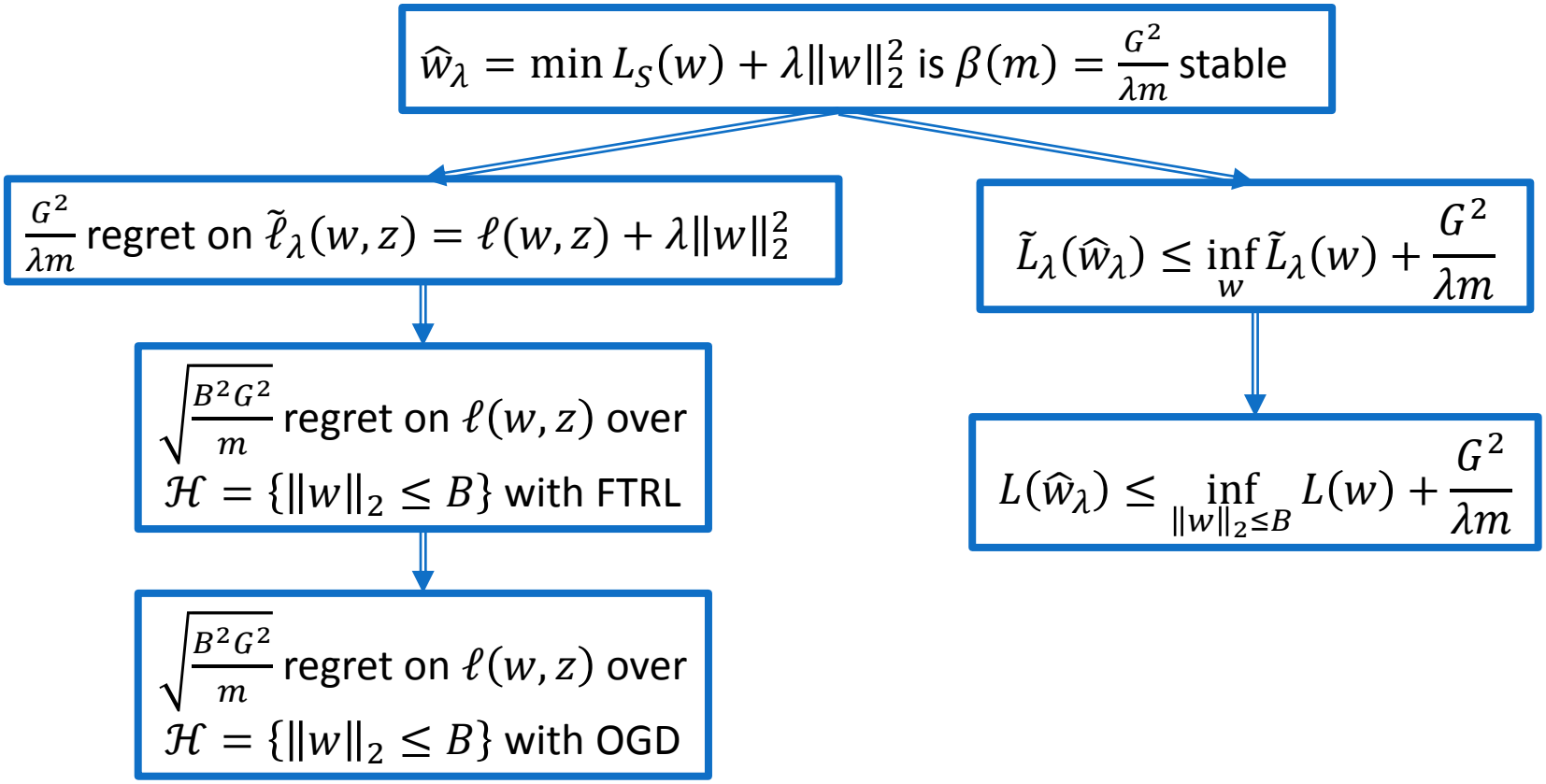
Prof. Nati Srebro

Lecture 15:

Strong Convexity

Online Dual Averaging and Mirror Descent

Stability and Learning



- So far: only convex Lipschitz bounded problems w.r.t. $\|w\|_2$
- **Today: Convex Lipschitz bounded problems w.r.t. other norms**
- Also: Cleaner online gradient stepsizes without scaling iterates

Convex Lipschitz Problems

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

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

- Until now: mostly concerned with $\overline{\mathcal{H}} = \mathcal{B} = \mathbb{R}^d$, i.e. improper learning when we allow any "predictor", and just want to compete with $\mathcal{H}$
- Will also consider $\overline{\mathcal{H}} \subset \mathcal{B}$, where we also insist predictor is in some restricted **convex** class $\overline{\mathcal{H}}$
  - $\overline{\mathcal{H}} = \mathcal{H}$, proper learning. E.g. we insist on finding a low-norm predictor
  - $\mathcal{H} \subset \overline{\mathcal{H}} \subset \mathcal{B}$, restricted improper learning
  - Might want to restrict $\overline{\mathcal{H}}$ to ensure Lipschitz inside $\overline{\mathcal{H}}$

Strong Convexity

- **Definition**: $\Psi: \overline{\mathcal{H}} \to \mathbb{R}$ is α-strongly convex w.r.t a norm $\|w\|$ if

$$\forall_{w,w' \in \overline{\mathcal{H}}} \Psi(w') \geq \Psi(w) + \langle \nabla \Psi(w), w' - w \rangle + \frac{\alpha}{2} \|w' - w\|^2$$

- **Claim**: if Ψ is α-strongly convex, and $w_0 = \arg\min_{w \in \overline{\mathcal{H}}} \Psi(w)$, then

$$\forall_{w \in \overline{\mathcal{H}}} \Psi(w) - \Psi(w_0) \geq \frac{\alpha}{2} \|w - w_0\|^2$$

- **Claim**: if Ψ is α-strongly convex, then $c\Psi$ is $(c \cdot \alpha)$-strongly convex
- **Claim**: if $f(w)$ is convex and $\Psi(w)$ is α-strongly convex, then $f(w) + \Psi(w)$ is α-strongly convex

- E.g. $\Psi(w) = \frac{1}{2} \|w\|_2^2$ is 1-strongly convex w.r.t $\|w\|_2$

Proof: $\frac{1}{2} \|w\|_2^2 + \langle w, w' - w \rangle + \frac{1}{2} \|w' - w\|_2^2 = \|w + (w' - w)\|_2^2 = \|w'\|^2$

Strong Convexity and Stability

- **Claim**: if $\ell(w,z)$ is G-Lipschitz w.r.t $\|w\|$ and $\Psi(w)$ is α-strongly convex w.r.t $\|w\|$, then

$$A(S) = \arg\min_{w\in\mathcal{H}} L_S(w) + \Psi(w)$$

is $\beta(m) \leq \frac{2G^2}{m\alpha}$ (leave-last-out or replacement) stable

Proof (leave-last-out):

- $f(w) = \frac{1}{m}\sum_{i=1}^{m} z_i(w) + \Psi(w)$, $w^* = \arg\min f(w)$ ➔ $f(w') - f(w^*) \geq \frac{\alpha}{2}\|w' - w^*\|^2$
- $f'(w) = \frac{1}{m}\sum_{i=1}^{m-1} z_i(w) + \frac{m-1}{m}\Psi(w)$, $w' = \arg\min f'(w)$

➔ $f'(w^*) - f'(w') \geq \frac{\alpha}{2}\frac{m-1}{m}\|w^* - w'\|^2$

➔ $\alpha\frac{2m-1}{2m}\|w^* - w'\|^2 \leq f'(w^*) - f(w^*) + f(w') - f'(w') = \frac{1}{m}\left(z_m(w') - z_m(w^*)\right)$
$\leq \frac{G}{m}\|w^* - w'\|$ ➔ $\|w^* - w'\| \leq \frac{2G}{\alpha(2m-1)}$ ➔ $|z_m(w^*) - z_m(w')| \leq \frac{2G^2}{\alpha(2m-1)}$

FTRL: Take II

$$w_{t+1} = FTRL(w_1, \dots, w_t) = \arg\min_{w \in \overline{\mathcal{H}}} \frac{1}{t} \sum_{i=1}^{t} \ell(w, z_i) + \lambda_t \Psi(w)$$

- For a convex G-Lipschitz problem w.r.t. norm $\|w\|$, if $\Psi(w) \geq 0$ is α-strongly convex w.r.t. $\|w\|$, for any $w \in \overline{\mathcal{H}}$,

$$\frac{1}{m} \sum_{t=1}^{m} \ell(w_t, z_t) \leq \frac{1}{\mathrm{m}} \sum_{t=1}^{m} \ell(w, z_t) + \frac{1}{m} \sum_{t=1}^{m} \lambda_t \, \Psi(w) + \frac{1}{m} \sum_{t=1}^{m} \frac{2G^2}{\alpha \lambda_t t}$$

- If also $\forall_{w \in \mathcal{H}} \Psi(w) \leq B^2$, using $\lambda_t = \sqrt{\frac{2G^2}{\alpha B^2 t}}$, for any $w \in \mathcal{H}$,

$$\frac{1}{m} \sum_{t=1}^{m} \ell(w_t, z_t) \leq \frac{1}{\mathrm{m}} \sum_{t=1}^{m} \ell(w, z_t) + \sqrt{\frac{32G^2B^2}{\alpha m}}$$

  i.e. FTRL has regret $\sqrt{\frac{32G^2B^2}{\alpha m}}$ relative to $\mathcal{H}$.

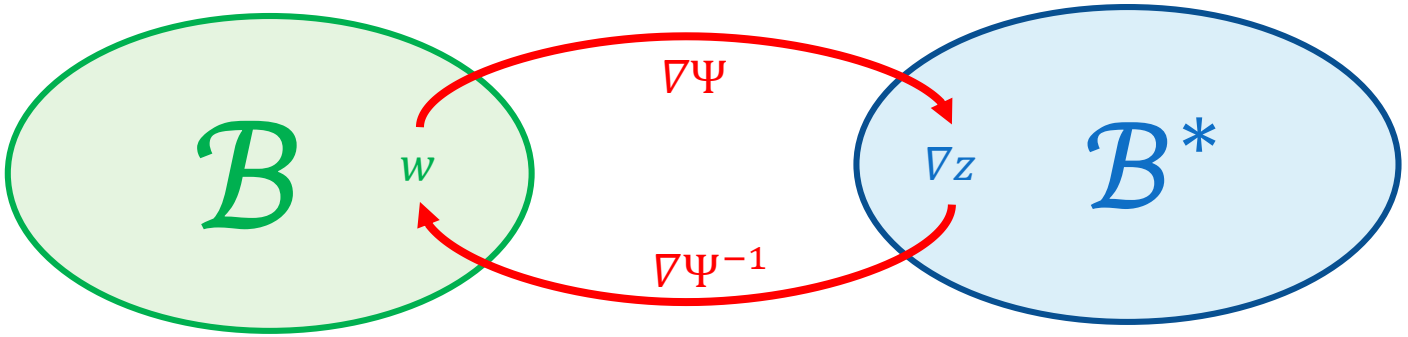
Linearized FTRL

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

- Same regret as FTRL!
- Only need to keep tack of averaged gradient $v_t = \frac{1}{t} \sum_{i=1}^{t} \nabla z_i(w_i)$
- If $\overline{\mathcal{H}} = \mathcal{B}$: $0 = \frac{1}{t} \sum_{i=1}^{t} \nabla z_i(w_i) + \lambda_t \nabla \Psi(w_{t+1})$

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

  ➔ $w_{t+1} = \nabla \Psi^{-1} \left(\frac{\lambda_{t-1}(t-1)}{\lambda_t t} \nabla \Psi(w_t) - \frac{1}{\lambda_t t} \nabla z_i(w_i) \right)$

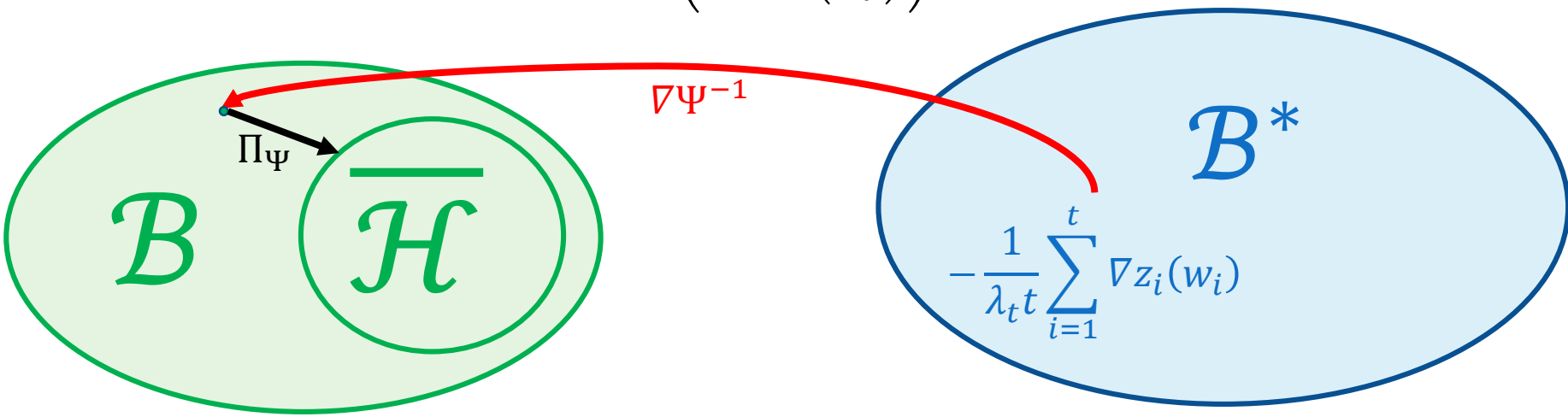


Linearized FTRL: "Dual Averaging"

- If $\overline{\mathcal{H}} = \mathcal{B}$: $w_{t+1} = \nabla\Psi^{-1}\left(-\frac{1}{\lambda_t t}\sum_{i=1}^{t}\nabla z_i(w_i)\right)$
- If $\overline{\mathcal{H}} \subset \mathcal{B}$: $w_{t+1} = \Pi_{\Psi}^{\overline{\mathcal{H}}}\left(\nabla\Psi^{-1}\left(-\frac{1}{\lambda_t t}\sum_{i=1}^{t}\nabla z_i(w_i)\right)\right)$
- Where: $\Pi_{\Psi}^{\mathcal{W}}(w) = \arg\min_{w'\in\mathcal{W}} D_{\Psi}(w'||w)$
- Bergman Divergence: $D_{\Psi}(w'||w) = \Psi(w') - (\Psi(w) + \langle\nabla\Psi(w), w' - w\rangle)$

Proof:

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



Example: $\|w\|_2$

- $\Psi(w) = \frac{1}{2}\|w\|_2^2$ is 1-strongly convex
- $\nabla\Psi(w) = w$, $\nabla\Psi^{-1}(\nu) = \nu$
- $D_\Psi(w'||w) = \Psi(w') - (\Psi(w) + \langle\nabla\Psi(w), w' - w\rangle)$

$$= \frac{1}{2}\|w'\|_2^2 - \left(\frac{1}{2}\|w\|_2^2 + \langle w, w' - w\rangle\right) + \|w\|_2^2 - \langle w, w\rangle$$

$$= \frac{1}{2}\|w - w'\|_2^2$$

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

- $\Psi(w) = \frac{1}{2} w^T Q w$ is 1-strongly convex w.r.t $\|w\|_Q$
- $\nabla\Psi(w) = Qw$, $\nabla\Psi^{-1}(\nu) = Q^{-1}\nu$
- $D_\Psi(w'||w) = \frac{1}{2}\|w - w'\|_Q^2$
- $\|\nu\|_* = \|\nu\|_{Q^{-1}} = \sqrt{\nu^T Q^{-1} \nu}$
- FTRL: $w_{t+1} = \frac{(t-1)\lambda_{t-1}}{t\lambda_t} w_t - Q^{-1}\nabla z_t(w_t)$
- Regret: $O\left(\sqrt{\frac{(w^T Q w)\left(\phi(x)^T Q^{-1}\phi(x)\right)}{m}}\right)$

Example: $\|w\|_p$

- $\Psi(w) = \frac{1}{2}\|w\|_p^2$ is $(p-1)$-strongly convex w.r.t. $\|w\|_p$
- $\nabla\Psi(w) = \|w\|_p^{2-p}|w[i]|^{p-1}sign(w[i])$
- $\nabla\Psi^{-1}(v) = \frac{|v[i]|^{q-1}sign(v[i])}{\|v\|_q^{q-2}}$, where $\frac{1}{p} + \frac{1}{q} = 1$
- Regret: $O\left(\sqrt{\frac{\|w\|_p^2\|\phi(x)\|_q^2}{(p-1)m}}\right)$
- Explodes as $p \to 1$

- What about $\|w\|_1$?
  - Option 1: use $q = \log d$ (Homework)
  - Option 2: entropy regularizer

$$\mathcal{H} = \overline{\mathcal{H}} = \{ w \mid w \geq 0, \|w\|_1 = 1 \}$$

- $\Psi(w) = \sum_i w[i] \log \frac{w[i]}{1/d} = \log d + \sum_i w[i] \log w[i]$
- 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 $\overline{\mathcal{H}}$
- $\nabla\Psi(w)[i] = \log w[i] + 1$
- $\nabla\Psi^{-1}(\nu)[i] = e^{\nu[i]-1}$

$$\nu_t = \frac{1}{t}\sum_{i=1}^{t} \nabla z_i(w_i)$$

- $w_{t+1} = \arg\min_{w \in \mathcal{H}} \langle \nu_t, w \rangle + \lambda_t \Psi(w)$ ➔ $w_{t+1}[i] = \frac{e^{-\frac{1}{\lambda_t}\nu_t[i]}}{\sum_i e^{-\frac{1}{\lambda_t}\nu_t[j]}}$
- Regret: $O\left(\sqrt{\frac{\|\phi(x)\|_\infty^2 \log d}{m}}\right)$

$$\mathcal{H} = \overline{\mathcal{H}} = \{ w \mid w \geq 0, \|w\|_1 = 1 \}$$

- $w_{t+1} = \arg\min_{w \in \mathcal{H}} \langle v_t, w \rangle + \lambda_t \Psi(w)$ ➔ $w_{t+1}[i] = \frac{e^{-\frac{1}{\lambda_t} v_t[i]}}{\sum_i e^{-\frac{1}{\lambda_t} v_t[j]}}$
- With $\lambda_t = \frac{\lambda}{t}$, we have $\frac{1}{\lambda_t} v_t = \frac{1}{\lambda} \sum_{i=1}^{t} \nabla z_i(w_i) = \frac{1}{\lambda_{t-1}} v_{t-1} + \frac{1}{\lambda} \nabla z_i(w_i)$
- The resulting updates are:

Normalized Exponentiated Gradient (EG)

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

- Recall regret: $O\left(\sqrt{\frac{\|\nabla z_t\|_\infty^2 \log d}{m}}\right)$

Back to Finite Cardinality

- Consider a finite cardinality hypothesis class $\mathcal{H}$ and bounded loss $0 \leq \ell \leq 1$ (e.g. $0/1$ error)
- HALVING: regret $\frac{\log|\mathcal{H}|}{m}$ in the realizable case
- What about agnostic case? Or general bounded loss?
- Solution: convexification
- Represent $\mathcal{H} = \{(1,0,\dots,0),(0,1,0,\dots,0),\dots,(0,0,\dots,0,1)\} \in \mathbb{R}^d$, where $d = |\mathcal{H}|$
- Linear loss, $\ell(w,z) = \langle w, \phi(z)\rangle$, where
$$\phi(z) = \left(\ell(h^1,z), \ell(h^2,z), \dots, \ell\left(h^{|\mathcal{H}|},z\right)\right) \in \mathbb{R}^d$$
is vector of losses each hypothesis in $\mathcal{H}$ would have on z
- $\mathcal{H}$ is non-convex, but we'll use improper learning with
$$\overline{\mathcal{H}} = \{\, w \in \mathbb{R}^d \mid \; w \geq 0, \|w\|_1 = 1\}$$
- Use normalized EG algorithm

Finite Cardinality Classes

Multiplicative Weights Algorithm

- $w_1[i] = \frac{1}{d}$

At round t:

- Pick hypothesis i w.p. $w_t[i]$
- $w_{t+1}[i] = \frac{w_t[i]e^{-\frac{1}{\lambda}\ell\left(h^i, z_t\right)}}{\sum_j w_t[j]e^{-\frac{1}{\lambda}\ell\left(h^j, z_t\right)}}$

Loss if using hypothesis i at round t

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

Nick Littlestone

From FTRL to Mirror Descent

- FTRL easily implementable only for linear objectives (or for a linearization)---**otherwise need to store all the history**
- Linearized FTRL ***almost*** yields online gradient descent:

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

but only **if** $\overline{\mathcal{H}} = \mathcal{B}$