Follow the Leader: Theory and Applications

Amit Kadan

Machine Learning Reading Group

July 3, 2019
Theory
Online Convex Optimization

**Input:** A convex set $S$.

for $t = 1, 2, ...$ :

- predict $w_t \in S$.
- receive convex loss $f_t : S \rightarrow \mathbb{R}$.
- suffer loss $f_t(w_t)$.

- Want to minimize regret with respect to best fixed action,

$$R(w_{1:T}) = \sum_{t=1}^{T} (f_t(w_t) - \min_{u \in S} f_t(u)) = \sum_{t=1}^{T} (f_t(w_t) - f_t(w^*))$$

- Specifically, want sub-linear regret $R(w_{1:T}) = o(T)$.

  - Implies average regret vanishes as $T \rightarrow \infty$.
  - Such algorithms known as no-regret algorithms.
Follow the Leader

- **Basic Idea**: Play strategy with minimal loss over past rounds.

Follow the Leader (FTL)

\[ w_t = \underset{w \in S}{\text{arg min}} \sum_{i=1}^{t-1} f_i(w). \]

- Is FTL a *no-regret* algorithm?
  - If optimize over losses up to *and including* loss at \( t \) expect to do well.
  - How much worse do we do such an algorithm?

Be the Leader (BTL)

\[ w_t = \underset{w \in S}{\text{arg min}} \sum_{i=1}^{t} f_i(w) \]

**Lemma [KV05]**

For any sequence of losses, BTL has non-positive regret.

**Corollary**

Let \( w_1, ..., w_t \) be iterates produced by FTL then,

\[ R(w_1:T) \leq \sum_{t=1}^{T}(f_t(w_t) - f_t(w_{t+1})) \]
FTL: Quadratic Loss

**Theorem [SS+12]**

Let \( S = \mathbb{R}^n \). Let \( f_t(w) = \frac{1}{2} \| w - z_t \|^2 \) for some \( z_t \in \mathbb{R}^n \). Then FTL suffers

\[
R(w_{1:T}) = O(\log T)
\]

Proof.

\( w_t \) has closed form solution, \( w_t = \frac{1}{t-1} \sum_{i=1}^{t-1} w_i \).

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

Use Corollary,

\[
f_t(w_t) - f_t(w_{t+1}) = \frac{1}{2} \| w_t - z_t \|^2 + \frac{1}{2} \| w_{t+1} - z_t \|^2
\]

\[
= \frac{1}{2} (1 - (1 - \frac{1}{t})^2) \| w_t - z_t \|^2
\]

\[
\leq \frac{1}{t} \| w_t - z_t \|^2
\]

Letting \( L = \max_t \| z_t \| \), since \( w_t \) is average of \( Z_i \), by triangle inequality,

\[
R_T(w_{1:T}) \leq (2L)^2 (\log T + 1) = o(T)
\]
FTL: Linear Loss

Example: Two Expert Setting [SS+12]

Let $S = [-1, 1]$. Let $f_t(w) = z_t w$. Define losses,

$$z_t = \begin{cases} 
-0.5 & t = 1 \\
1 & t \text{ even} \\
-1 & t \text{ odd}
\end{cases}$$

Suffer loss at least $T - 1$ at time $T$, best expert has loss $T/2$. Regret is $T/2 - 1 = O(T)$.

- What causes FTL to do poorly on linear losses?
- Easy fixes to allow FTL no-regret for general convex $f_t$’s?
For rest of presentation, we focus on linear loss, i.e., for some \( z_t \in \mathbb{R}^n \), \( f_t(w) = \langle w, z_t \rangle \).

**Note:** Expert setting is specific instance of linear losses where restrict \( w_t \) to lie in the simplex, i.e., \( w_t^j \leq 1 \) and \( \sum_j w_t^j = 1 \), and \( z_t \in [0, 1]^n \).
Stabilization

- **Notice:** for two experts, \( w_t \)'s unstable (oscillate between -1, 1).
- **Solution:** add stability.

Follow the Regularized Leader (FTRL)

Let \( L : S \to \mathbb{R} \) be strongly-convex regularizer, FTRL updates are given,
\[
 w_t = \arg \min_{w \in S} \sum_{i=1}^{t-1} \langle w, z_i \rangle + L(w)
\]

- Bound regret of FTRL relative to Be the Regularized Leader (BTRL)

**Lemma \([SS^{+}12]\)**

Regret of FTRL is bounded by BTRL loss,
\[
 R(w_1:T) \leq \sum_{t=1}^{T} (f_t(w_t) - f_t(w_{t+1})) + L(w^*) - L(w_1)
 = \sum_{t=1}^{T} \langle w_t - w_{t+1}, z_t \rangle + L(w^*) - L(w_1)
\]
FTRL with L2 regularization (GD update)

Consider FTRL on $S = \mathbb{R}^n$ with $L(w) = \frac{1}{2\eta} \|w\|^2_2$,

$$w_t = \arg\min_{w \in \mathbb{R}^n} \sum_{i=1}^{t-1} \langle z_i, w \rangle + \frac{1}{2\eta} \|w\|^2_2$$

Taking the derivative, and solving for $w$,

$$w_t = -\eta \sum_{t=1} w_t - \eta z_{t-1}$$

If optimizing a single convex function $f$ with loss vector $\nabla f(w_t)$, get usual gradient descent update,

$$w_t = w_{t-1} - \eta \nabla f(w_{t-1})$$
FTRL: Regularization Schemes

FTRL with entropy regularization (MWU update)

Consider FTRL for expert setting with entropy regularizer

\[ L(w) = \frac{1}{\eta} \sum_j w^j \log w^j \quad (1\text{-strongly-convex in L1 norm}). \]

FTRL updates given by,

\[
    w_t = \arg \min_{w \in S} \sum_{i=1}^{t-1} \langle w, z_i \rangle + \frac{1}{\eta} \sum_j w^j \log w^j
\]

If write first-order Lagrangian with constraint \( \sum_j w^j = 1 \) get,

\[
    L(w, \lambda) = \sum_{i=1}^{t-1} \langle z_i, w \rangle + \frac{1}{\eta} \sum_j w^j \log w^j + \lambda(1 - \sum_j w^j)
\]

Solving, we get multiplicative weight update,

\[
    w^k_t = \frac{\exp(-\eta \sum_{i=1}^{t-1} z^k_i)}{\sum_j \exp(-\eta \sum_{i=1}^{t-1} z^j_i)}
\]
FTRL: Linear Loss

**Theorem [SS+12]**

Consider running FTRL with regularizer $L(w) = \frac{1}{2\eta} \|w\|_2^2$. Assume $\forall w \in S$, $\|w\|_2 \leq B$, and assume $\|z_t\|_2 \leq C$. Choosing $\eta = \frac{B}{C \sqrt{2T}}$, we have $R(w_{1:T}) = O(\sqrt{T})$.

Proof.
By Lemma,

$$R(w_{1:T}) \leq L(w^*) - L(w_1) + \sum_{t=1}^{T} \langle w_t - w_{t+1}, z_t \rangle$$

$$\leq \frac{1}{2\eta} \|w^*\|_2^2 + \sum_{t=1}^{T} \langle w_t - w_{t+1}, z_t \rangle$$

$$= \frac{B^2}{2\eta} + \eta \sum_{t=1}^{T} \|z_t\|_2^2 \quad \text{(GD update)}$$

$$\leq \frac{B^2}{2\eta} + \eta TC^2 = \frac{B}{C} \sqrt{\frac{2}{T}}$$

In fact, FTRL no-regret when $L(w)$ 1-strongly convex in norm defined on $S$. 
Randomization

- **Alternative interpretation of FTL failure:** Poor synchronization with losses.
- **Solution:** Add randomness to predictor sequence.

Follow the Perturbed Leader (FTPL) proposed by Kalai and Vempala [KV05]. FTPL uses FTL with extra hallucinated cost at time 0 sampled from distribution.

Follow the Perturbed Leader (FTPL) [KV05]

Let $z_0 \sim d\mu(x)$. Then FTPL updates given by,

$$w_t = \arg \min_{w \in S} \sum_{i=0}^{t-1} \langle w, z_i \rangle$$
FTPL: Linear Loss

Theorem [KV05]

Let \( D = \sup_{x,y \in S} \| x - y \|_1 \), \( A = \sup_{1 \leq t \leq T} \| z_t \| \).

Define probability distribution \( d\mu(x) = \left( \frac{\epsilon}{2} \right)^n e^{-\epsilon \| x \|_1} \).

FTPL with \( z_0 \) sampled from \( d\mu(x) \) satisfies,

\[
\mathbb{E} \left[ \sum_{t=1}^{T} \langle z_t, w_t \rangle \right] \leq O(1 + \epsilon A) \mathbb{E}[\langle z_t, u \rangle] + O\left( \frac{D}{\epsilon} \log n \right)
\]

- Proof similar to FTRL proof.
- Above bound contains both additive and multiplicative constant.
- FTPL can be used to solve problems where number of experts is exponential in size of input.
- FTPL can also be shown to be no-regret (only additive constant) with \( R(w_{1:T}) = O(\sqrt{T}) \), with a modified version of \( d\mu(x) \) [KV05].
Applications
GAN Training

- During the training of a Generative Adversarial Network (GAN), a discriminator and a generator compete in a two player game.
- GANs often suffer from lack of convergence during training.
- GANs have generated an interest in analyzing two player game dynamics, i.e., settings where both players use learning algorithms to try and converge to some sort of equilibrium.
- No-regret learning algorithms are thought to be good candidates for such dynamics.
A well-known result due to Robinson [Rob51] characterizes convergence of game where both players use FTL.

Fictitious Play [Rob51]

Consider following two-player game. Let $A \in \mathbb{R}^{m \times n}$, consider payoff function $\phi(x, y) = x^T Ay$, i.e., player 1 suffers loss $x^T Ay$, and player 2 suffers loss $-x^T Ay$. Assume player 1 uses learning rule,

$$x_t = \arg \min_{j} \sum_{i=1}^{t-1} \langle e_j, Ay_i \rangle$$

And player 2 uses learning rule,

$$y_t = \arg \max_{j} \sum_{i=1}^{t-1} \langle e_j, A^T x_i \rangle$$

Then players’ plays converge to a Nash equilibrium of game.
Robinson’s result revived by Ge et al. [GXC+18] to propose Fictitious-GAN - a training algorithm based on fictitious play.

Show when discriminator and generator use fictitious play as opposed to gradient descent/ascent, convergence in the players’ utilities.

(a) Fictitious play

(b) Gradient descent/ascent dynamics
Zhen and Kwok [ZK17] develop Follow the Moving Leader (FTML), a novel optimization algorithm for training deep networks.

Use an FTL variant, Follow the Proximal Regularized Leader (FTPRL) update rule.

**Follow the Proximal Regularized Leader (FTPRL)**

\[
w_t = \arg \min_{w \in S} \sum_{i=1}^{t} P_i(w) = \arg \min_{w \in S} \sum_{i=1}^{t} (\langle g_i, w \rangle + \frac{1}{2} \| w - w_{i-1} \|_{Q_i}^2)
\]

Some of the most well-known algorithms for deep network optimization such as Adagrad, Adam, can be derived as FTPRL update.
Zhen and Kwok [ZK17] use FTPRL algorithm to compare FTML to other state-of-the-art optimization algorithms.

For instance, FTML uses a weighted FTPRL update,

\[ w_t = \arg\min_{w \in S} \sum_{i=1}^{t} w_{i,t} (\langle g_i, w \rangle + \frac{1}{2} \| w - w_{i-1} \|_Q^2) \]

Adam can be written in a very similar way except that rather then centering each \( P_i(w) \) at \( w_{i-1} \) it centers them all at \( w_{t-1} \),

\[ w_t = \arg\min_{w \in S} \sum_{i=1}^{t} w_{i,t} (\langle g_i, w \rangle + \frac{1}{2} \| w - w_{t-1} \|_Q^2) \]

Suggest that centring on only last iterate results in Adam being less stable than FTML in changing environments.
In experiments, Zheng and Kwok [ZK17] show that FTML outperforms state-of-the-art optimization algorithms such as Adam, RMSProp, and Adadelta on various deep learning objectives.
Recap

Theory

- Follow the Leader is a natural algorithm for online learning.
- FTL is not a no-regret algorithm.
- Two perspectives to turn into no-regret algorithm.
  1. Regularization: FTRL.
  2. Randomization: FTPL.

Applications

- Game theoretic perspective + no-regret algorithms can result in novel techniques for training GANs (notoriously difficult to train).
- Can view many state-of-the-art optimization algorithms as variants of FTL, devise new methods, compare through FTL setup.
- Other applications include dual averaging techniques for convex optimization.


