Acceleration through Optimistic No-Regret Dynamics

Jun-Kun Wang and Jacob Abernethy

Georgia Tech

Convex Optimization

$$\min_{x \in \mathcal{X}} f(x) \tag{1}$$

Method: Gradient Descent, Frank-Wolfe method, Nesterov's accelerated method, Heavy Ball ... etc.

Convex Optimization

$$\min_{x \in \mathcal{X}} f(x) \tag{1}$$

Method: Gradient Descent, Frank-Wolfe method, Nesterov's accelerated method, Heavy Ball ... etc.

L-smooth convex problems $\min_{x \in \mathcal{X}} f(x)$.

- : Nesterov's accelerated method: $O(\frac{1}{T^2})$.

L-smooth and μ-strongly convex problems $\min_{x \in \mathcal{X}} f(x)$. Denote $\kappa := \frac{L}{\mu}$.

- : Nesterov's accelerated method: $O(\exp(-\frac{T}{\sqrt{\kappa}}))$.

Online learning (minimizing regret)

Online learning protocol:

1: **for** $t = 1$ to T **do**
2: Play w_t according to $\mathit{OnlineAlgorithm}^w\big(\ell_1(w_1), \ldots, \ell_{t-1}(w_{t-1})\big)$.
3: Receive loss function $\ell_t(\cdot)$ and suffer loss $\ell_t(w_t)$.
4: **end for**

$\text{Regret}_T^w := \sum_{t=1}^T \ell_t(w_t) - \sum_{t=1}^T \ell_t(w^*)$.

convex loss functions $\{\ell_t(\cdot)\}_{t=1}^T$.

- $\frac{\text{Regret}_T^w}{T} = O(\frac{1}{\sqrt{T}})$.

strongly convex loss functions $\{\ell_t(\cdot)\}_{t=1}^T$.

- $\frac{\text{Regret}_T^w}{T} = O(\frac{\log T}{T})$.

New perspective: A two-player zero-sum game

A zero-sum game (*Fenchel game*)

$$g(x, y) := \langle x, y \rangle - f^*(y).$$

$$V^* := \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} g(x, y) \overset{def}{=} \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \langle x, y \rangle - f^*(y) \overset{Fenchel}{=} \min_{x \in \mathcal{X}} f(x).$$

New perspective: A two-player zero-sum game

A zero-sum game (*Fenchel game*)

$g(x, y) := \langle x, y \rangle - f^*(y).$

$$V^* := \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} g(x, y) \overset{def}{=} \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \langle x, y \rangle - f^*(y) \overset{Fenchel}{=} \min_{x \in \mathcal{X}} f(x).$$

Equivalent to solving the underlying optimization problem!

If $(\hat{x}, \hat{y})$ is an ϵ-equilibrium of the game, then

$$f(\hat{x}) \leq \min_x f(x) + \epsilon.$$

Meta algorithm for Fenchel-game

Algorithm 0 Meta Algorithm

1: Given a sequence of weights $\{\alpha_t\}$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: $\quad y_t := \mathit{OnlineAlgorithm}^Y(g(x_1, \cdot), \ldots, g(x_{t-1}, \cdot)$.
4: $\quad x_t := \mathit{OnlineAlgorithm}^X(g(\cdot, y_1), \ldots, g(\cdot, y_{t-1}), g(\cdot, y_t))$.
5: $\quad$ y-player's loss function: $\alpha_t \ell_t(y) := \alpha_t(f^*(y) - \langle x_t, y \rangle)$.
6: $\quad$ x-player's loss function: $\alpha_t h_t(x) := \alpha_t(\langle x, y_t \rangle - f^*(y_t))$.
7: **end for**
8: Output $(\bar{x}_T, \bar{y}_T) := \left(\frac{\sum_{s=1}^T \alpha_s x_s}{A_T}, \frac{\sum_{s=1}^T \alpha_s y_s}{A_T} \right)$.

Let $x^* = \arg\min_x f(x)$.

$$\alpha\text{-}\mathrm{REG}^y := \textstyle\sum_{t=1}^T \alpha_t \ell_t(y_t) - \min_{y \in \mathcal{Y}} \sum_{t=1}^T \alpha_t \ell_t(y) \tag{2}$$

$$\alpha\text{-}\mathrm{REG}^x := \sum_{t=1}^T \alpha_t h_t(x_t) - \sum_{t=1}^T \alpha_t h_t(x^*) \tag{3}$$

Meta algorithm for Fenchel-game

Algorithm 0 Meta Algorithm

1: Given a sequence of weights $\{\alpha_t\}$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: $\quad y_t := OnlineAlgorithm^Y(g(x_1, \cdot), \ldots, g(x_{t-1}, \cdot)$.
4: $\quad x_t := OnlineAlgorithm^X(g(\cdot, y_1), \ldots, g(\cdot, y_{t-1}), g(\cdot, y_t))$.
5: $\quad$ y-player's loss function: $\alpha_t \ell_t(y) := \alpha_t(f^*(y) - \langle x_t, y \rangle)$.
6: $\quad$ x-player's loss function: $\alpha_t h_t(x) := \alpha_t(\langle x, y_t \rangle - f^*(y_t))$.
7: **end for**
8: Output $(\bar{x}_T, \bar{y}_T) := \left(\frac{\sum_{s=1}^T \alpha_s x_s}{A_T}, \frac{\sum_{s=1}^T \alpha_s y_s}{A_T} \right)$.

Define the weighted average regret $\overline{\boldsymbol{\alpha}\text{-}\mathrm{REG}} := \frac{\boldsymbol{\alpha}\text{-}\mathrm{REG}}{A_T}$, $A_T := \sum_{t=1}^T \alpha_t$.

Theorem: $f(\bar{x}_T) \leq \min_x f(x) + \frac{\boldsymbol{\alpha}\text{-}\mathrm{REG}^x}{A_T} + \frac{\boldsymbol{\alpha}\text{-}\mathrm{REG}^y}{A_T}$.

(Unconstrained Optimization: $\min_{x \in \mathbb{R}^n} f(x)$))

Algorithm 1 Nesterov's method from the Meta Algorithm

1: Given the sequence of weights $\{\alpha_t = t\}$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: $\quad$ y-player plays `Optimistic-FTL`.
$\quad$ $y_t \leftarrow \nabla f(\widetilde{x}_t) = \arg\min_{y \in \mathcal{Y}} \sum_{s=1}^{t-1} \alpha_s \ell_s(y) + m_t(y)$,
$\quad$ where $m_t(y) = \alpha_t \ell_{t-1}(y)$ and $\widetilde{x}_t := \frac{1}{A_t}(\alpha_t x_{t-1} + \sum_{s=1}^{t-1} \alpha_s x_s)$.
4: $\quad$ x-player plays `Gradient Descent`.
5: $\quad$ $x_t = x_{t-1} - \gamma_t \alpha_t \nabla h_t(x) = x_{t-1} - \gamma_t \alpha_t y_t = x_{t-1} - \gamma_t \alpha_t \nabla f(\widetilde{x}_t)$.
6: **end for**
7: Output $(\bar{x}_T, \bar{y}_T) := \left(\frac{\sum_{s=1}^{T} \alpha_s x_s}{A_T}, \frac{\sum_{s=1}^{T} \alpha_s y_s}{A_T} \right)$.

$$\bar{x}_{t+1} = \bar{x}_t - \frac{1}{4L} \nabla f(\widetilde{x}_{t+1}) + (\tfrac{t-1}{t+2})(\bar{x}_t - \bar{x}_{t-1}).$$

Other accelerated variants

(Constrained Optimization: $\min_{x \in \mathcal{K}} f(x)$))

Algorithm 2 Nesterov's method from the Meta Algorithm

1: Given the sequence of weights $\{\alpha_t = t\}$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: y-player plays `Optimistic-FTL`.
$y_t \leftarrow \nabla f(\widetilde{x}_t) = \arg\min_{y \in \mathcal{Y}} \sum_{s=1}^{t-1} \alpha_s \ell_s(y) + m_t(y)$,
where $m_t(y) = \alpha_t \ell_{t-1}(y)$ and $\widetilde{x}_t := \frac{1}{A_t}(\alpha_t x_{t-1} + \sum_{s=1}^{t-1} \alpha_s x_s)$.
4: (A) x-player plays `Mirror Descent`.
5: $x_t = \arg\min_{x \in \mathcal{K}} \gamma_t \langle x, \alpha_t y_t \rangle + V_{x_{t-1}}(x)$.
6: Or, (B) x-player plays `Be-The-Regularized-Leader`.
7: $x_t = \arg\min_{x \in \mathcal{K}} \sum_{s=1}^{t} \theta_s \langle x, \alpha_s y_s \rangle + \frac{1}{\eta} R(x)$,
8: **end for**
9: Output $(\bar{x}_T, \bar{y}_T) := \left(\frac{\sum_{s=1}^{T} \alpha_s x_s}{A_T}, \frac{\sum_{s=1}^{T} \alpha_s y_s}{A_T} \right)$.

(A) Nesterov's 1988 (1-memory) and (B) Nesterov's 2005 (∞-memory) accelerated method

Heavy Ball method

(Unconstrained Optimization: $\min_{x \in \mathbb{R}^n} f(x)$))

Algorithm 3 Heavy Ball from the Meta Algorithm

1: Given the sequence of weights $\{\alpha_t = t\}$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: $\quad$ y-player plays `FTL` .
$\quad$ $y_t \leftarrow \nabla f(\bar{x}_{t-1}) = \arg\min_{y \in \mathcal{Y}} \sum_{s=1}^{t-1} \alpha_s \ell_s(y)$ $\bar{x}_{t-1} := \frac{\sum_{s=1}^{t-1} \alpha_s x_s}{A_{t-1}}$
4: $\quad$ x-player plays `Gradient Descent` .
5: $\quad$ $x_t = x_{t-1} - \gamma_t \alpha_t \nabla h_t(x) = x_{t-1} - \gamma_t \alpha_t y_t = x_{t-1} - \gamma_t \alpha_t \nabla f(\widetilde{x}_t)$.
6: **end for**
7: Output $(\bar{x}_T, \bar{y}_T) := \left(\frac{\sum_{s=1}^{T} \alpha_s x_s}{A_T}, \frac{\sum_{s=1}^{T} \alpha_s y_s}{A_T} \right)$.

$$\bar{x}_t = \bar{x}_{t-1} - \frac{\gamma_t \alpha_t^2}{A_t} \nabla f(\bar{x}_{t-1}) + (\frac{\alpha_t A_{t-2}}{A_t \alpha_{t-1}})(\bar{x}_{t-1} - \bar{x}_{t-2}). \text{ (Heavy ball)}$$

$$\bar{x}_t = \bar{x}_{t-1} - \frac{\gamma_t \alpha_t^2}{A_t} \nabla f(\widetilde{x}_t) + (\frac{\alpha_t A_{t-2}}{A_t \alpha_{t-1}})(\bar{x}_{t-1} - \bar{x}_{t-2}). \text{ (Nesterov's alg.)}$$

Analysis: L-smooth convex optimization problems

y-player plays `Optimistic-FTL`

$y_t \leftarrow \nabla f(\widetilde{x}_t) = \arg\min_{y \in \mathcal{Y}} \sum_{s=1}^{t-1} \alpha_s \ell_s(y) + \alpha_t \ell_{t-1}(y)$

$$\alpha\text{-}\mathrm{REG}^y := \sum_{t=1}^{T} \alpha_t \ell_t(y_t) - \min_{y \in \mathcal{Y}} \sum_{t=1}^{T} \alpha_t \ell_t(y) \leq \sum_{t=1}^{T} \frac{L\alpha_t^2}{A_t} \|x_{t-1} - x_t\|^2.$$

x-player plays `MirrorDescent`

$x_t = \arg\min_{x \in \mathcal{K}} \gamma_t' \langle \nabla f(\widetilde{x}_t), x \rangle + V_{x_{t-1}}(x)$

$$\alpha\text{-}\mathrm{REG}^x := \sum_{t=1}^{T} \alpha_t h_t(x_t) - \sum_{t=1}^{T} \alpha_t h_t(x^*) \leq \frac{D}{\gamma_T} - \sum_{t=1}^{T} \frac{1}{2\gamma_t} \|x_{t-1} - x_t\|^2.$$

where D is a constant such that $V_{x_t}(x^*) \leq D$ for all t.

$$f(\bar{x}_T) - \min_{x \in \mathcal{X}} f(x) \leq \frac{1}{A_T} \left(\frac{D}{\gamma_T} + \sum_{t=1}^{T} \cancel{\left(\frac{\alpha_t^2}{A_t} L - \frac{1}{2\gamma_t}\right)\|x_{t-1} - x_t\|^2} \right) = O\left(\frac{LD}{T^2}\right).$$

as long as γ_t satisfying $\frac{1}{CL} \leq \gamma_t \leq \frac{1}{4L}$ for some constant $C > 4$.

Thank you!

Other instances of the meta-algorithm

- Accelerated linear rate of Nesterov's method for strongly convex and smooth problems
- Accelerated Proximal Method
- Accelerated Frank-Wolfe

Come to our poster #156!