MetaGrad: Faster Convergence Without Curvature in Online Convex Optimization

Wouter M. Koolen  Tim van Erven

Inria Lille
Friday 15th April, 2016
This talk

- Online Convex Optimization
- Learning the Learning rate
- Second-order (variance) bounds (individual sequence)
- Fast rates without curvature
Outline

Online Convex Optimization

A New Type of Guarantee

Fast Rates

MetaGrad Algorithm
Fundamental Learning Model: Online Convex Optimization

► In round \( t = 1, 2, \ldots \)
  ► Learner predicts \( \mathbf{w}_t \) (from unit ball)
  ► Encounter convex loss function \( f_t(\mathbf{u}) : \mathbb{R}^d \to \mathbb{R} \)

► Learner
  ► observes gradient \( \mathbf{g}_t := \nabla f_t(\mathbf{w}_t) \) (from unit ball)
  ► incurs loss \( f_t(\mathbf{w}_t) \)
In round $t = 1, 2, \ldots$

- Learner predicts $w_t$ (from unit ball)
- Encounter convex loss function $f_t(u) : \mathbb{R}^d \to \mathbb{R}$

Learner

- observes gradient $g_t := \nabla f_t(w_t)$ (from unit ball)
- incurs loss $f_t(w_t)$

The goal is to have small regret

$$R_T^u := \sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(u)$$

with respect to every point $u$. 
The Learner’s Perspective

Round 1: Learner plays $w_1 = 0$
The Learner’s Perspective

Round 1: Learner incurs $f_1(w_1)$ and sees $g_1 = \nabla f_1(w_1)$
The Learner’s Perspective

Round 2: Learner plays $w_1 = 1/4$
The Learner’s Perspective

Round 2: Learner incurs $f_2(w_2)$ and sees $g_2 = \nabla f_2(w_2)$
The Learner’s Perspective
Evaluate Learner using regret: \( R_T^u = \sum_{t=1}^{T} (f_t(w_t) - f_t(u)) \)
State of the Art

Online gradient descent

\[ w_{t+1} = w_t - \eta g_t \]

recall \( g_t = \nabla f_t(w_t) \)
Online gradient descent

\[ w_{t+1} = w_t - \eta g_t \]

recall \( g_t = \nabla f_t(w_t) \)

OGD bound: After \( T \) rounds,

\[
R^u_T \leq O \left( \sqrt{\sum_{t=1}^{T} \| g_t \|^2} \right)
\]

for all \( u \) with \( \| u \| \leq 1 \).
Always have **worst-case guarantee**

\[
R_T^u \leq O \left( \sqrt{\sum_{t=1}^{T} \|g_t\|^2} \right) \leq O(\sqrt{T}).
\]

Yet bound says we **might get lucky**

For smooth functions \(f_t\) with common optimum \(u^*\), as \(w_t \to u^*\), we have \(g_t \to 0\), and

\[
\sqrt{\sum_{t=1}^{T} \|g_t\|^2} \ll \sqrt{T}
\]

**grows much slower** than \(\sqrt{T}\).
What We Hope Happens

\[ R_T \leq O \left( \sqrt{\sum_{t=1}^{T} \| g_t \|^2} \right) \]
Can We Do Better?

No in general: matching lower bound.

\[ R_u^T \geq \Omega(\sqrt{T}) \]

Yes, with curvature:

\[ R_u^T \leq O(\ln T) \]

- Strongly convex: \( I \preceq \nabla^2 f(u) \), e.g.
  \[ f_t(u) = \|u - y_t\|^2 \]
  \( \Rightarrow \) gradient descent with small \( \eta \)

- Exp-concave: \( \nabla f(u)\nabla f(u)^\top \preceq \nabla^2 f(u) \), e.g.
  \[ f_t(u) = -\ln (1 + y_t^\top u) \]
  \( \Rightarrow \) Online Newton Step
Can We Do Better?

But do we really need \textit{curvature}? 

This talk: no, \textit{stability} is enough.

New algorithm MetaGrad: Separate learning rate $\eta$ for each point $u$
Outline

Online Convex Optimization

A New Type of Guarantee

Fast Rates

MetaGrad Algorithm
Refined Bound

Recall bound for gradient descent:

\[ R_T^u \leq O \left( \sqrt{\sum_{t=1}^{T} \| g_t \|^2} \right) \]

New bound for MetaGrad:

\[ R_T^u \leq O \left( \sqrt{V_T^u d \ln T} \right) \quad \text{where} \quad V_T^u := \sum_{t=1}^{T} \left( (w_t - u)^T g_t \right)^2 \]

Data-dependent. Whoa! Ouroboric.

Always improvement:

\[ \left( (w_t - u)^T g_t \right)^2 \leq \| w_t - u \|^2 \| g_t \|^2 \]
Now What We Hope Happens

\[ R_T^u \leq O \left( \sqrt{\sum_{t=1}^{T} ((w_t - u)^\top g_t)^2} \right) \]
Outline

Online Convex Optimization

A New Type of Guarantee

Fast Rates

MetaGrad Algorithm
Does it Really Work?

Offline optimization (fixed function):

\[ f_t(u) = |u - 1/4| \]
The “Fast Rates” Pipeline

Combine

- refined individual-sequence regret bound

\[ R_T^u \leq \sqrt{V_T^u d \ln T} \quad \forall u \]

- Special-purpose argument that for best \( u^* \)

\[ V_T^{u^*} \leq R_T^{u^*} \]

- Profit!

\[ R_T^{u^*} \leq \sqrt{R_T^{u^*} d \ln T} \quad \text{so} \quad R_T^{u^*} \leq d \ln T \]
Any fixed $f_t(u) = f(u)$.
Let $u^* = \arg \min_u f(u)$ be the offline minimiser.

Crux: $(w_t - u^*)^T g_t \in [0, 2]$.

Now from the regret bound

$$R_{T}^{u^*} \leq \sum_{t=1}^{T} (w_t - u^*)^T g_t \leq \sqrt{V_{T}^{u^*} d \ln T}$$

and special-purpose observation

$$V_{T}^{u^*} = \sum_{t=1}^{T} ((u^* - w_t)^T g_t)^2 \leq 2 \sum_{t=1}^{T} (w_t - u^*)^T g_t$$

we can solve for $V_{T}^{u^*}$ to find $V_{T}^{u^*} \leq d \ln T$ and hence

$$R_{T}^{u^*} \leq \sqrt{2d \ln T}$$
Does It Really Actually Work?

Stochastic optimization:

\[ f_t(u) = |u - x_t| \]

where \( x_t = \pm \frac{1}{2} \) i.i.d. with probability 0.4 and 0.6.
What’s Going On, Really?

Stochastic optimization:

\[ f_t(u) = |u - x_t| \]

where \( x_t = \pm \frac{1}{2} \) i.i.d. with probability 0.4 and 0.6.

Stable minimum easy to converge to
Significant Improvement: Stochastic Case

Consider i.i.d.

\[ f \sim \mathbb{P} \quad \text{with} \quad u^* = \arg \min_u \mathbb{E}[f(u)] \]

Condition: there is a \( c > 0 \) such that

\[
\forall w : (w - u^*)^\top \mathbb{E} [\nabla f(w) \nabla f(w)^\top] (w - u^*) \leq c (w - u^*)^\top \mathbb{E} [\nabla f(w)]
\]

Now from the special-case condition

\[
\mathbb{E}[V_T^{u^*}] \leq c \mathbb{E} \left[ \sum_{t=1}^T (w_t - u^*)^\top \nabla f(w_t) \right]
\]

and by the generic regret bound, in expectation,

\[
\mathbb{E}[R_T^{u^*}] \leq \mathbb{E} \left[ \sum_{t=1}^T (w_t - u^*)^\top \nabla f(w_t) \right] \leq \mathbb{E} \left[ \sqrt{V_T^{u^*} d \ln T} \right]
\]

and by Jensen’s inequality \( \mathbb{E} \left[ \sqrt{V_T^{u^*}} \right] \leq \sqrt{\mathbb{E} \left[ V_T^{u^*} \right]} \), so that

\[
\mathbb{E}[R_T^{u^*}] \leq \sqrt{cd \ln T}
\]
Outline

Online Convex Optimization

A New Type of Guarantee

Fast Rates

MetaGrad Algorithm
Our Approach in a Nutshell

1. Replace actual loss $f_t(u)$ by **surrogate loss** $\ell^n_t(u)$
   - parametrised by learning rate $\eta$
   - exp-concave in $u$
   - So can get good bound for surrogate regret

2. Exponentially spaced **grid** $\eta_1, \eta_2, \ldots, \eta_{\log(T)}$ ($\eta_i = 2^{-i}$).

3. Off-the-shelf exp-concave **Slave** for grid point $\eta_i$ predicts $w^{\eta_i}_1, w^{\eta_i}_2, \ldots$

4. At each round $t$, **Master** aggregates $w^{\eta_1}_t, w^{\eta_2}_t, \ldots$ into $w_t$. 


**Surrogate Loss**

Real loss

\[ f_t(u) \leq f_t(w_t) + (u - w_t)^T g_t \]

Surrogate loss

\[ \ell_t^\eta(u) \ := \ \eta(u - w_t)^T g_t + (\eta(u - w_t)^T g_t)^2 \]

Exp-concave! In particular:

\[ e^{-\ell_t^\eta(u)} \leq 1 + \eta(w_t - u)^T g_t. \]

Excellent bound \( O(\ln T) \) for wrong loss.
MetaGrad Slave

$\eta$-Slave (variant of Online Newton Step) predicts

$$w_{t+1}^\eta = w_t^\eta - \eta \Sigma_{t+1}^\eta g_t$$

where the covariance matrix is given by

$$\Sigma_{t+1}^\eta = \left(\frac{1}{4}I + 2\eta^2 \sum_{s=1}^{t} g_s g_s^\top\right)^{-1}$$

$\eta$-Slave guarantees

$$\sum_{t=1}^{T} (\ell_t^\eta(w_t^\eta) - \ell_t^\eta(u)) \leq \frac{1}{8}\|u\|^2 + \frac{1}{2} \ln \det \left(I + 8\eta^2 \sum_{t=1}^{T} g_t g_t^\top\right)$$

$$\leq O(d \ln T) \quad \forall u$$
Input: Grid points $\eta_i = 2^{-i}$ with weights $\pi_i = \frac{1}{i(i+1)}$.
Goal: aggregate $w_{\eta_1^t}, w_{\eta_2^t}, \ldots$

Idea: Potential

$$\Phi_t := \sum_i \pi_i e^{-\sum_{s=1}^t \ell_s^{\eta_i}(w_s^{\eta_i})}.$$ 

Two steps:

- Find predictions $w_t$ that ensure $1 \geq \Phi_1 \geq \Phi_2 \geq \ldots$
- Derive regret bound from $1 \geq \Phi_T$. 
MetaGrad Master, Potential Decreases

Tilted exponentially weighted average

\[
\mathbf{w}_{t+1} = \frac{\sum_i \pi_i e^{-\sum_{s=1}^{t} \ell_{s}^{\eta_i}(\mathbf{w}_{s}^{\eta_i})} \eta_i \mathbf{w}_{t+1}}{\sum_i \pi_i e^{-\sum_{s=1}^{t} \ell_{s}^{\eta_i}(\mathbf{w}_{s}^{\eta_i})} \eta_i}
\]

ensures potential shrinks:

\[
\Phi_{t+1} - \Phi_t = \sum_i \pi_i e^{-\sum_{s=1}^{t} \ell_{s}^{\eta_i}(\mathbf{w}_{s}^{\eta_i})} \left( e^{-\ell_{t+1}^{\eta_i}(\mathbf{w}_{t+1})} - 1 \right)
\]

\[
\exp\text{-con} \leq \sum_i \pi_i e^{-\sum_{s=1}^{t} \ell_{s}^{\eta_i}(\mathbf{w}_{s}^{\eta_i})} \eta_i (\mathbf{w}_{t+1} - \mathbf{w}_{t+1}^{\eta_i})^\top \mathbf{g}_{t+1} \quad \text{weights} = 0
\]

and hence \(\Phi_t \leq 1\).
The Master achieves for all $t$:

$$1 \geq \Phi_t = \sum_i \pi_i e^{-\sum_{s=1}^{t} \ell_{s_i}^{\eta_i}(w_{s_i}^{\eta_i})}.$$

It follows that

$$\sum_{t=1}^{T} (0 - \ell_{t_i}^{\eta_i}(w_{t_i}^{\eta_i})) \leq -\ln \pi_i \quad \forall i \text{ in grid}$$

(Master has zero surrogate loss)
MetaGrad Analysis

Now combine the Master and Slave guarantee. For each grid point \( \eta \) and comparator \( u \)

\[
\sum_{t=1}^{T} (0 - \ell^\eta_t(w^\eta_t)) \leq - \ln \pi_i \leq \ln \ln T
\]

\[
\sum_{t=1}^{T} (\ell^\eta_t(w^\eta_t) - \ell^\eta_t(u)) \leq O(d \ln T)
\]

so

\[
\sum_{t=1}^{T} (0 - \ell^\eta_t(u)) \leq O(d \ln T).
\]

Unpacking \( \ell^\eta_t(u) = \eta(u - w_t)^T g_t + (\eta(u - w_t)^T g_t)^2 \) yields

\[
\eta R^u_T \leq \eta^2 V^u_T + O(d \ln T).
\]
MetaGrad Analysis (ctd.)

Reorganise the bound to:

$$R^u_T \leq \eta V^u_T + \frac{O(d \ln T)}{\eta}$$

Now pick the best grid point

$$\hat{\eta} = \sqrt{\frac{O(d \ln T)}{V^u_T}}$$

to find

$$R^u_T \leq O\left(\sqrt{V^u_T d \ln T}\right)$$

of course we need a grid point close to $\hat{\eta}$ and we need to deal with off-grid $\hat{\eta} \gg 1$ and $\hat{\eta} \ll \frac{1}{\sqrt{T}}$. 
MetaGrad Outlook

- Run-time $O(d^2)$ per round
- Projections (avoid $O(d^3)$ per round!)
- We design and analyze two versions of Slave
  - Full covariance (quadratic)
  - Diagonal approximation (linear)
- Very welcome to discuss further
MetaGrad Outlook

- Run-time $O(d^2)$ per round
- Projections (avoid $O(d^3)$ per round!)
- We design and analyze two versions of Slave
  - Full covariance (quadratic)
  - Diagonal approximation (linear)
- Very welcome to discuss further

Learn more:
- Paper submitted to COLT 2016, preprint available
- Code is available
  http://bitbucket.org/wmkoolen/metagrad
- Experiments coming soon.
  http://blog.wouterkoolen.info
Summary

Low regret through stability, even without curvature.

- New MetaGrad algorithm.
- Hierarchical Master-Slave construction.
- Learns the learning rate.
- Refined (adaptive) regret bound.
- Stochastic condition for logarithmic regret (fast rates)
Thank you!