Leiden Lorentz Workshop, November 9, 2016

MetaGrad: Multiple Learning Rates in Online Learning

Tim van Erven Wouter Koolen

Universiteit Leiden

CWI Centrum Wiskunde & Informatica

Online Convex Optimization

Parameters $\boldsymbol{w}$ take values in a convex domain $\mathcal{U}$

for $t = 1, 2, \ldots, T$ **do**
    Learner plays $\boldsymbol{w}_t \in \mathcal{U}$
    Environment reveals convex loss function $f_t : \mathcal{U} \to \mathbb{R}$
    Learner incurs loss $f_t(\boldsymbol{w}_t)$, observes gradient $\boldsymbol{g}_t = \nabla f_t(\boldsymbol{w}_t)$
end for

Measure **regret** w.r.t. $u \in \mathcal{U}$:

$$\text{Regret}_T^{\boldsymbol{u}} = \sum_{t=1}^{T} f_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} f_t(\boldsymbol{u})$$

Assumptions: diameter$(\mathcal{U}) \leq D$, $\|\boldsymbol{g}_t\|_2 \leq G$.

The Standard Picture

Rates based on curvature:

Convex f_t	$\sqrt{T}$	GD with $\eta_t \propto \frac{1}{\sqrt{t}}$
Strongly convex f_t	$\ln T$	GD with $\eta_t \propto \frac{1}{t}$
Exp-concave f_t	$d \ln T$	ONS with $\eta_t =$ constant

- [Bartlett, Hazan, and Rakhlin, 2007], [Do et al., 2009]: Adaptive GD: **strongly convex + general convex**

The Standard Picture

Rates based on curvature:

Convex f_t	$\sqrt{T}$	GD with $\eta_t \propto \frac{1}{\sqrt{t}}$
Strongly convex f_t	$\ln T$	GD with $\eta_t \propto \frac{1}{t}$
Exp-concave f_t	$d \ln T$	ONS with $\eta_t =$ constant

- [Bartlett, Hazan, and Rakhlin, 2007], [Do et al., 2009]: Adaptive GD: **strongly convex + general convex**

Our goals:

- Adaptivity to more types of functions f_t
- Fast rates **without curvature** for 'easy' stochastic data

Other Types of Adaptivity

- [Orabona, 2014, Orabona and Pál, 2016]:
  adapt to size $\|\boldsymbol{u}\|_2$ of comparator
- AdaGrad [Duchi et al., 2011]:
  box-like domain (ℓ_∞-ball) instead of ℓ_2-ball
- [Hazan and Kale, 2010], [Chiang et al., 2012]:
  linear functions f_t that vary little over time
- [Orabona, Crammer, and Cesa-Bianchi, 2015]:
  data-dependent time-varying regularizers
- ...

Key techniques:

- Adaptive tuning of learning rate η_t
- Use second-order information about covariance of features in time-varying regularizer

MetaGrad: Multiple Eta Gradient Algorithm

Theorem

MetaGrad's $\text{Regret}_T^{\boldsymbol{u}}$ is bounded by

$$\sum_{t=1}^{T} (\boldsymbol{w}_t - \boldsymbol{u})^\intercal \boldsymbol{g}_t = \begin{cases} O\left(\sqrt{V_T^{\boldsymbol{u}}\, d \ln T} + d \ln T\right) \\ \\ O(\sqrt{T \ln \ln T}), \end{cases}$$

where

$$V_T^{\boldsymbol{u}} = \sum_{t=1}^{T} (\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t \boldsymbol{g}_t^\intercal (\boldsymbol{u} - \boldsymbol{w}_t).$$

- By convexity, $f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{u}) \leq (\boldsymbol{w}_t - \boldsymbol{u})^\intercal \boldsymbol{g}_t$.

MetaGrad: Multiple Eta Gradient Algorithm

Theorem

MetaGrad's $\text{Regret}_T^{\boldsymbol{u}}$ is bounded by

$$\sum_{t=1}^{T} (\boldsymbol{w}_t - \boldsymbol{u})^\intercal \boldsymbol{g}_t = \begin{cases} O\left(\sqrt{V_T^{\boldsymbol{u}}\, d \ln T} + d \ln T\right) \\ \\ O(\sqrt{T \ln \ln T}), \end{cases}$$

where

$$V_T^{\boldsymbol{u}} = \sum_{t=1}^{T} (\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t \boldsymbol{g}_t^\intercal (\boldsymbol{u} - \boldsymbol{w}_t).$$

- By convexity, $f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{u}) \leq (\boldsymbol{w}_t - \boldsymbol{u})^\intercal \boldsymbol{g}_t$.
- Covariance: $\boldsymbol{g}_t \boldsymbol{g}_t^\intercal \propto \boldsymbol{X}_t \boldsymbol{X}_t^\intercal$ when $f_t(\boldsymbol{w}) = \text{loss}(Y_t \langle \boldsymbol{w}, \boldsymbol{X}_t \rangle)$

MetaGrad: Multiple Eta Gradient Algorithm

Theorem

MetaGrad's $\text{Regret}_T^{\boldsymbol{u}}$ is bounded by

$$\sum_{t=1}^{T} (\boldsymbol{w}_t - \boldsymbol{u})^\intercal \boldsymbol{g}_t = \begin{cases} O\left(\sqrt{V_T^{\boldsymbol{u}}\, d \ln T} + d \ln T\right) \\ \\ O(\sqrt{T \ln \ln T}), \end{cases}$$

where

$$V_T^{\boldsymbol{u}} = \sum_{t=1}^{T} (\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t \boldsymbol{g}_t^\intercal (\boldsymbol{u} - \boldsymbol{w}_t).$$

- By convexity, $f_t(\boldsymbol{w}_t) - f_t(\boldsymbol{u}) \leq (\boldsymbol{w}_t - \boldsymbol{u})^\intercal \boldsymbol{g}_t$.
- Covariance: $\boldsymbol{g}_t \boldsymbol{g}_t^\intercal \propto \boldsymbol{X}_t \boldsymbol{X}_t^\intercal$ when $f_t(\boldsymbol{w}) = \text{loss}(Y_t \langle \boldsymbol{w}, \boldsymbol{X}_t \rangle)$
- Optimal learning rate depends on $V_T^{\boldsymbol{u}}$, but $\boldsymbol{u}$ unknown!
  Solution: aggregate **multiple learning rates**.

Consequences

Non-stochastic adaptation:

Convex f_t	$\sqrt{T \ln \ln T}$
Exp-concave f_t	$d \ln T$
Fixed convex $f_t = f$	$d \ln T$

Consequences

Non-stochastic adaptation:

Convex f_t	$\sqrt{T \ln\ln T}$
Exp-concave f_t	$d \ln T$
Fixed convex $f_t = f$	$d \ln T$

Loose end: strongly convex $\Rightarrow$ exp-concave gives $d \ln T$

Consequences

Non-stochastic adaptation:

Convex f_t	$\sqrt{T \ln\ln T}$
Exp-concave f_t	$d \ln T$
Fixed convex $f_t = f$	$d \ln T$

Loose end: strongly convex $\Rightarrow$ exp-concave gives $d \ln T$

Stochastic without curvature [joint work with Grünwald]

Suppose f_t i.i.d. with stochastic optimum $\boldsymbol{u}^* = \arg\min_{\boldsymbol{u} \in \mathcal{U}} \mathbb{E}_f[f(\boldsymbol{u})]$. Then expected regret $\mathbb{E}[\text{Regret}_T^{\boldsymbol{u}^*}]$:

Absolute loss* $f_t(u) = \lvert u - X_t \rvert$	$\ln T$
Hinge loss* $\max\{0, 1 - Y_t \langle \boldsymbol{u}, \boldsymbol{X}_t \rangle\}$	$d \ln T$
(B, β)-Bernstein	$(Bd \ln T)^{1/(2-\beta)}\, T^{(1-\beta)/(2-\beta)}$

*Conditions apply

Outline

Two General Fast Rate Conditions

MetaGrad Algorithm

Exponential Weights Interpretation of Online Learning Algorithms

1. Directional Derivative Condition

Theorem

If there exist $a, b > 0$ such that all f_t satisfy

$$f_t(\boldsymbol{u}) \geq f_t(\boldsymbol{w}) + a(\boldsymbol{u} - \boldsymbol{w})^\intercal \nabla f_t(\boldsymbol{w}) + b\left((\boldsymbol{u} - \boldsymbol{w})^\intercal \nabla f_t(\boldsymbol{w})\right)^2 \quad \text{for } \boldsymbol{w} \in \mathcal{U},$$

then $O(d \ln T)$ regret w.r.t. $\boldsymbol{u}$.

$a = 1$

- Satisfied by **exp-concave** functions [Hazan, Agarwal, and Kale, 2007]
- Requires quadratic curvature in direction of minimizer $\boldsymbol{u}$.

General a

- Satisfied for any **fixed convex** function $f_t = f$ with minimizer $\boldsymbol{u}$, even without any curvature, with $a = 2$ and $b = 1/(DG)$.

2. Bernstein Condition for Online Learning

Suppose f_t i.i.d. with stochastic optimum $\boldsymbol{u}^* = \operatorname*{arg\,min}_{\boldsymbol{u} \in \mathcal{U}} \mathbb{E}_f[f(\boldsymbol{u})]$.

Standard Bernstein condition:

$$\mathbb{E}\,(f(\boldsymbol{w}) - f(\boldsymbol{u}^*))^2 \leq B\Big(\mathbb{E}\,[f(\boldsymbol{w}) - f(\boldsymbol{u}^*)]\Big)^\beta \qquad \text{for all } \boldsymbol{w} \in \mathcal{U}.$$

2. Bernstein Condition for Online Learning

Suppose f_t i.i.d. with stochastic optimum $\boldsymbol{u}^* = \arg\min_{\boldsymbol{u} \in \mathcal{U}} \mathbb{E}_f[f(\boldsymbol{u})]$.

Standard Bernstein condition:

$$\mathbb{E}\left(f(\boldsymbol{w}) - f(\boldsymbol{u}^*)\right)^2 \leq B\big(\mathbb{E}\left[f(\boldsymbol{w}) - f(\boldsymbol{u}^*)\right]\big)^\beta \qquad \text{for all } \boldsymbol{w} \in \mathcal{U}.$$

Replace by **weaker linearized version:**

- Apply with $\tilde{f}(\boldsymbol{u}) = \langle \boldsymbol{u}, \nabla f(\boldsymbol{w}) \rangle$ instead of f!
- By convexity, $f(\boldsymbol{w}) - f(\boldsymbol{u}^*) \leq \tilde{f}(\boldsymbol{w}) - \tilde{f}(\boldsymbol{u}^*)$.

$$\mathbb{E}\left((\boldsymbol{w} - \boldsymbol{u}^*)\nabla f(\boldsymbol{w})\right)^2 \leq B\big(\mathbb{E}\left[(\boldsymbol{w} - \boldsymbol{u}^*)\nabla f(\boldsymbol{w})\right]\big)^\beta \qquad \text{for all } \boldsymbol{w} \in \mathcal{U}.$$

2. Bernstein Condition for Online Learning

Suppose f_t i.i.d. with stochastic optimum $\boldsymbol{u}^* = \arg\min_{\boldsymbol{u} \in \mathcal{U}} \mathbb{E}_f[f(\boldsymbol{u})]$.

Standard Bernstein condition:

$$\mathbb{E}\left(f(\boldsymbol{w}) - f(\boldsymbol{u}^*)\right)^2 \le B\Big(\mathbb{E}\left[f(\boldsymbol{w}) - f(\boldsymbol{u}^*)\right]\Big)^\beta \qquad \text{for all } \boldsymbol{w} \in \mathcal{U}.$$

Replace by **weaker linearized version:**

- Apply with $\tilde{f}(\boldsymbol{u}) = \langle \boldsymbol{u}, \nabla f(\boldsymbol{w}) \rangle$ instead of f!
- By convexity, $f(\boldsymbol{w}) - f(\boldsymbol{u}^*) \le \tilde{f}(\boldsymbol{w}) - \tilde{f}(\boldsymbol{u}^*)$.

$$\mathbb{E}\left((\boldsymbol{w} - \boldsymbol{u}^*)\nabla f(\boldsymbol{w})\right)^2 \le B\Big(\mathbb{E}\left[(\boldsymbol{w} - \boldsymbol{u}^*)\nabla f(\boldsymbol{w})\right]\Big)^\beta \qquad \text{for all } \boldsymbol{w} \in \mathcal{U}.$$

Hinge loss (with $G = D = 1$): $\beta = 1$, $B = \frac{2\lambda_{\max}(\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\intercal])}{\|\mathbb{E}[Y\boldsymbol{X}]\|}$

2. Bernstein Condition for Online Learning

Suppose f_t i.i.d. with stochastic optimum $\boldsymbol{u}^* = \arg\min_{\boldsymbol{u}\in\mathcal{U}} \mathbb{E}_f[f(\boldsymbol{u})]$.

Standard Bernstein condition:

$$\mathbb{E}\left(f(\boldsymbol{w}) - f(\boldsymbol{u}^*)\right)^2 \leq B\Big(\mathbb{E}\left[f(\boldsymbol{w}) - f(\boldsymbol{u}^*)\right]\Big)^\beta \qquad \text{for all } \boldsymbol{w} \in \mathcal{U}.$$

Replace by **weaker linearized version:**

- Apply with $\tilde{f}(\boldsymbol{u}) = \langle \boldsymbol{u}, \nabla f(\boldsymbol{w})\rangle$ instead of f!
- By convexity, $f(\boldsymbol{w}) - f(\boldsymbol{u}^*) \leq \tilde{f}(\boldsymbol{w}) - \tilde{f}(\boldsymbol{u}^*)$.

$$\mathbb{E}\left((\boldsymbol{w} - \boldsymbol{u}^*)\nabla f(\boldsymbol{w})\right)^2 \leq B\Big(\mathbb{E}\left[(\boldsymbol{w} - \boldsymbol{u}^*)\nabla f(\boldsymbol{w})\right]\Big)^\beta \qquad \text{for all } \boldsymbol{w} \in \mathcal{U}.$$

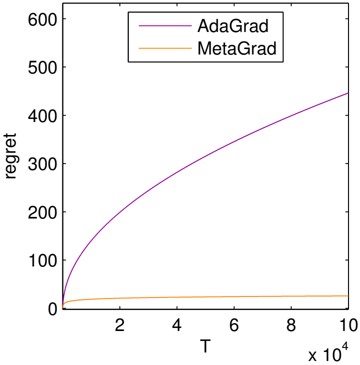
Hinge loss (with $G = D = 1$): $\beta = 1$, $B = \frac{2\lambda_{\max}(\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\intercal])}{\|\mathbb{E}[Y\boldsymbol{X}]\|}$

Theorem (Koolen, Grünwald, Van Erven, 2016)

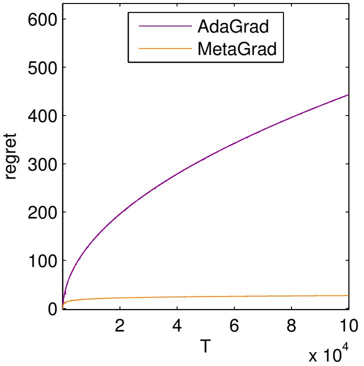
$$\mathbb{E}[\text{Regret}_T^{\boldsymbol{u}^*}] = O\Big((Bd\ln T)^{1/(2-\beta)}\, T^{(1-\beta)/(2-\beta)}\Big)$$

$$\text{Regret}_T^{\boldsymbol{u}^*} = O\Big((Bd\ln T - \ln\delta)^{1/(2-\beta)}\, T^{(1-\beta)/(2-\beta)}\Big) \quad w.p. \geq 1-\delta$$

Difference in Rates Not Just Theoretical



Offline: $f_t(u) = |u - 1/4|$



Stochastic Online: $f_t(u) = |u - X_t|$
where $X_t = \pm\frac{1}{2}$ i.i.d. w.p. 0.4 and 0.6.

- MetaGrad: $O(\ln T)$ regret, AdaGrad: $O(\sqrt{T})$, match bounds
- Functions neither strongly convex nor smooth
- **Caveat:** comparison more complicated for higher dimensions, unless we run a separate copy of MetaGrad per dimension, like the diagonal version of AdaGrad runs GD per dimension

Outline

Two General Fast Rate Conditions

MetaGrad Algorithm

Exponential Weights Interpretation of Online Learning Algorithms

MetaGrad Algorithm

Second-order **surrogate loss** for each η of interest (from a grid):

$$\ell_t^\eta(\boldsymbol{u}) = \eta(\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t + \eta^2 (\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t \boldsymbol{g}_t^\intercal (\boldsymbol{u} - \boldsymbol{w}_t)$$

MetaGrad Algorithm

Second-order **surrogate loss** for each η of interest (from a grid):

$$\ell_t^\eta(\boldsymbol{u}) = \eta(\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t + \eta^2 (\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t \boldsymbol{g}_t^\intercal (\boldsymbol{u} - \boldsymbol{w}_t)$$

One **Slave** algorithm per η produces $\boldsymbol{w}_t^\eta$ such that

$$\sum_{t=1}^{T} \ell_t^\eta(\boldsymbol{w}_t^\eta) - \sum_{t=1}^{T} \ell_t^\eta(\boldsymbol{u}) \leq R_{\text{slave}}^{\boldsymbol{u}}(\eta)$$

MetaGrad Algorithm

Second-order **surrogate loss** for each η of interest (from a grid):

$$\ell_t^\eta(\boldsymbol{u}) = \eta(\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t + \eta^2 (\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t \boldsymbol{g}_t^\intercal (\boldsymbol{u} - \boldsymbol{w}_t)$$

One **Slave** algorithm per η produces $\boldsymbol{w}_t^\eta$ such that

$$\sum_{t=1}^T \ell_t^\eta(\boldsymbol{w}_t^\eta) - \sum_{t=1}^T \ell_t^\eta(\boldsymbol{u}) \leq R_{\text{slave}}^{\boldsymbol{u}}(\eta)$$

Single **Master** algorithm produces $\boldsymbol{w}_t$ such that

$$\underbrace{\sum_{t=1}^T \ell_t^\eta(\boldsymbol{w}_t)}_{=0} - \sum_{t=1}^T \ell_t^\eta(\boldsymbol{w}_t^\eta) \leq R_{\text{master}}(\eta) \qquad \forall \eta$$

MetaGrad Algorithm

Second-order **surrogate loss** for each η of interest (from a grid):

$$\ell_t^\eta(\boldsymbol{u}) = \eta(\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t + \eta^2(\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t \boldsymbol{g}_t^\intercal (\boldsymbol{u} - \boldsymbol{w}_t)$$

One **Slave** algorithm per η produces $\boldsymbol{w}_t^\eta$ such that

$$\sum_{t=1}^T \ell_t^\eta(\boldsymbol{w}_t^\eta) - \sum_{t=1}^T \ell_t^\eta(\boldsymbol{u}) \le R_{\text{slave}}^{\boldsymbol{u}}(\eta)$$

Single **Master** algorithm produces $\boldsymbol{w}_t$ such that

$$\underbrace{\sum_{t=1}^T \ell_t^\eta(\boldsymbol{w}_t)}_{=0} - \sum_{t=1}^T \ell_t^\eta(\boldsymbol{w}_t^\eta) \le R_{\text{master}}(\eta) \qquad \forall \eta$$

Together: $-\sum_{t=1}^T \ell_t^\eta(\boldsymbol{u}) \le R_{\text{slave}}^{\boldsymbol{u}}(\eta) + R_{\text{master}}(\eta) \quad \forall \eta$

MetaGrad Algorithm

Second-order **surrogate loss** for each η of interest (from a grid):

$$\ell_t^\eta(\boldsymbol{u}) = \eta(\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t + \eta^2 (\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t \boldsymbol{g}_t^\intercal (\boldsymbol{u} - \boldsymbol{w}_t)$$

One **Slave** algorithm per η produces $\boldsymbol{w}_t^\eta$ such that

$$\sum_{t=1}^T \ell_t^\eta(\boldsymbol{w}_t^\eta) - \sum_{t=1}^T \ell_t^\eta(\boldsymbol{u}) \le R_{\text{slave}}^{\boldsymbol{u}}(\eta)$$

Single **Master** algorithm produces $\boldsymbol{w}_t$ such that

$$\underbrace{\sum_{t=1}^T \ell_t^\eta(\boldsymbol{w}_t)}_{=0} - \sum_{t=1}^T \ell_t^\eta(\boldsymbol{w}_t^\eta) \le R_{\text{master}}(\eta) \qquad \forall \eta$$

Together: $-\sum_{t=1}^T \ell_t^\eta(\boldsymbol{u}) \le R_{\text{slave}}^{\boldsymbol{u}}(\eta) + R_{\text{master}}(\eta) \quad \forall \eta$

$$\sum_{t=1}^T (\boldsymbol{w}_t - \boldsymbol{u})^\intercal \boldsymbol{g}_t \le \frac{R_{\text{slave}}^{\boldsymbol{u}}(\eta) + R_{\text{master}}(\eta)}{\eta} + \eta V_T^{\boldsymbol{u}}$$

MetaGrad Algorithm

Second-order **surrogate loss** for each η of interest (from a grid):

$$\ell_t^\eta(\boldsymbol{u}) = \eta(\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t + \eta^2(\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t \boldsymbol{g}_t^\intercal (\boldsymbol{u} - \boldsymbol{w}_t)$$

One **Slave** algorithm per η produces $\boldsymbol{w}_t^\eta$ such that

$$\sum_{t=1}^{T} \ell_t^\eta(\boldsymbol{w}_t^\eta) - \sum_{t=1}^{T} \ell_t^\eta(\boldsymbol{u}) \le R_{\text{slave}}^{\boldsymbol{u}}(\eta)$$

Single **Master** algorithm produces $\boldsymbol{w}_t$ such that

$$\underbrace{\sum_{t=1}^{T} \ell_t^\eta(\boldsymbol{w}_t)}_{=0} - \sum_{t=1}^{T} \ell_t^\eta(\boldsymbol{w}_t^\eta) \le R_{\text{master}}(\eta) \qquad \forall \eta$$

Together: $-\sum_{t=1}^{T} \ell_t^\eta(\boldsymbol{u}) \le R_{\text{slave}}^{\boldsymbol{u}}(\eta) + R_{\text{master}}(\eta) \quad \forall \eta$

$$\sum_{t=1}^{T} (\boldsymbol{w}_t - \boldsymbol{u})^\intercal \boldsymbol{g}_t \le \frac{O(d \ln T) + O(\ln \ln T)}{\eta} + \eta V_T^{\boldsymbol{u}}$$

MetaGrad Algorithm

Second-order **surrogate loss** for each η of interest (from a grid):

$$\ell_t^\eta(\boldsymbol{u}) = \eta(\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t + \eta^2(\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t \boldsymbol{g}_t^\intercal (\boldsymbol{u} - \boldsymbol{w}_t)$$

One **Slave** algorithm per η produces $\boldsymbol{w}_t^\eta$ such that

$$\sum_{t=1}^T \ell_t^\eta(\boldsymbol{w}_t^\eta) - \sum_{t=1}^T \ell_t^\eta(\boldsymbol{u}) \leq R_{\text{slave}}^{\boldsymbol{u}}(\eta)$$

Single **Master** algorithm produces $\boldsymbol{w}_t$ such that

$$\underbrace{\sum_{t=1}^T \ell_t^\eta(\boldsymbol{w}_t)}_{=0} - \sum_{t=1}^T \ell_t^\eta(\boldsymbol{w}_t^\eta) \leq R_{\text{master}}(\eta) \qquad \forall \eta$$

Together: $-\sum_{t=1}^T \ell_t^\eta(\boldsymbol{u}) \leq R_{\text{slave}}^{\boldsymbol{u}}(\eta) + R_{\text{master}}(\eta) \quad \forall \eta$

$$\sum_{t=1}^T (\boldsymbol{w}_t - \boldsymbol{u})^\intercal \boldsymbol{g}_t \leq \frac{O(d \ln T) + O(\ln \ln T)}{\eta} + \eta V_T^{\boldsymbol{u}} \Rightarrow O\left(\sqrt{V_T^{\boldsymbol{u}} d \ln T}\right)$$

MetaGrad Master

Goal: aggregate slave predictions $\boldsymbol{w}_t^\eta$ for all η in exponentially spaced grid $\frac{2^{-0}}{5DG}, \frac{2^{-1}}{5DG}, \ldots, \frac{2^{-\lceil \frac{1}{2}\log_2 T\rceil}}{5DG}$

Difficulty: master's predictions must be good w.r.t. different loss functions ℓ_t^η for all η simultaneously

Compute **exponential weights** with performance of each η measured by its own surrogate loss:

$$\pi_t(\eta) = \frac{\pi_1(\eta) e^{-\sum_{s<t} \ell_s^\eta(\boldsymbol{w}_s^\eta)}}{Z}$$

Then predict with **tilted** exponentially weighted average:

$$\boldsymbol{w}_t = \frac{\sum_\eta \pi_t(\eta)\, \eta\, \boldsymbol{w}_t^\eta}{\sum_\eta \pi_t(\eta)\, \eta}$$

MetaGrad Master Analysis

Potential $$\Phi_T = \sum_{\eta} \pi_1(\eta) e^{-\sum_{t=1}^{T} \ell_t^\eta(\boldsymbol{w}_t^\eta)}$$

Proof outline:

$$\Phi_T \leq \Phi_{T-1} \leq \cdots \leq \Phi_0 = 1$$

$$\pi_1(\eta) e^{-\sum_{t=1}^{T} \ell_t^\eta(\boldsymbol{w}_t^\eta)} \leq 1 \qquad \forall \eta$$

$$\underbrace{\sum_{t=1}^{T} \ell_t^\eta(\boldsymbol{w}_t)}_{=0} - \sum_{t=1}^{T} \ell_t^\eta(\boldsymbol{w}_t^\eta) \leq -\ln \pi_1(\eta)$$

MetaGrad Master Analysis

Potential $$\Phi_T = \sum_\eta \pi_1(\eta) e^{-\sum_{t=1}^T \ell_t^\eta(\boldsymbol{w}_t^\eta)}$$

Proof outline:

$$\Phi_T \le \Phi_{T-1} \le \cdots \le \Phi_0 = 1$$

$$\pi_1(\eta) e^{-\sum_{t=1}^T \ell_t^\eta(\boldsymbol{w}_t^\eta)} \le 1 \qquad \forall \eta$$

$$\underbrace{\sum_{t=1}^T \ell_t^\eta(\boldsymbol{w}_t)}_{=0} - \sum_{t=1}^T \ell_t^\eta(\boldsymbol{w}_t^\eta) \le -\ln \pi_1(\eta)$$

Grid has $\lceil \frac{1}{2} \log_2 T \rceil + 1$ learning rates, so for heavy-tailed prior:

$$-\ln \pi_1(\eta) = O(\ln \ln T)$$

MetaGrad Master Analysis: Decreasing Potential

Surrogate loss $\ell_t^\eta(\boldsymbol{u}) = \eta(\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t + \eta^2(\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t \boldsymbol{g}_t^\intercal (\boldsymbol{u} - \boldsymbol{w}_t)$ is **exp-concave**, even if f_t is not.

Upper bound by tangent at $\boldsymbol{u} = \boldsymbol{w}_t$:

$$e^{-\ell_t^\eta(\boldsymbol{u})} \leq 1 + \eta(\boldsymbol{w}_t - \boldsymbol{u})^\intercal \boldsymbol{g}_t$$

MetaGrad Master Analysis: Decreasing Potential

Surrogate loss $\ell_t^\eta(\boldsymbol{u}) = \eta(\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t + \eta^2(\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t \boldsymbol{g}_t^\intercal (\boldsymbol{u} - \boldsymbol{w}_t)$ is **exp-concave**, even if f_t is not.

Upper bound by tangent at $\boldsymbol{u} = \boldsymbol{w}_t$:

$$e^{-\ell_t^\eta(\boldsymbol{u})} \leq 1 + \eta(\boldsymbol{w}_t - \boldsymbol{u})^\intercal \boldsymbol{g}_t$$

Choose master's weights to ensure decreasing potential:

$$\begin{aligned}
\Phi_T - \Phi_{T-1} &= \sum_\eta \pi_1(\eta) e^{-\sum_{t<T} \ell_t^\eta(\boldsymbol{w}_t^\eta)} \left(e^{-\ell_T^\eta(\boldsymbol{w}_T^\eta)} - 1 \right) \\
&\leq \sum_\eta \pi_1(\eta) e^{-\sum_{t<T} \ell_t^\eta(\boldsymbol{w}_t^\eta)} \eta(\boldsymbol{w}_T - \boldsymbol{w}_T^\eta)^\intercal \boldsymbol{g}_T \\
&= 0 \qquad \text{for any } \boldsymbol{g}_T
\end{aligned}$$

MetaGrad Slave

Goal: Given η, minimize regret w.r.t. exp-concave surrogate ℓ_t^η.

Update:

$$\tilde{\boldsymbol{w}}_{t+1}^\eta = \boldsymbol{w}_t^\eta - \eta s_t^\eta \boldsymbol{\Sigma}_{t+1}^\eta \boldsymbol{g}_t,$$

where

$$\Sigma_{t+1}^\eta = \left(\frac{1}{D^2}\boldsymbol{I} + 2\eta^2 \sum_{s=1}^{t} \boldsymbol{g}_s \boldsymbol{g}_s^\intercal\right)^{-1} \qquad s_t^\eta = 1 + 2\eta \boldsymbol{g}_t^\intercal(\boldsymbol{w}_t^\eta - \boldsymbol{w}_t)$$

Project onto domain:

$$\boldsymbol{w}_{t+1}^\eta = \operatorname*{arg\,min}_{\boldsymbol{u}\in\mathcal{U}} \quad (\boldsymbol{u} - \tilde{\boldsymbol{w}}_{t+1}^\eta)^\intercal (\Sigma_{t+1}^\eta)^{-1} (\boldsymbol{u} - \tilde{\boldsymbol{w}}_{t+1}^\eta)$$

MetaGrad Slave

Goal: Given η, minimize regret w.r.t. exp-concave surrogate ℓ_t^η.

Update:

$$\tilde{\boldsymbol{w}}_{t+1}^\eta = \boldsymbol{w}_t^\eta - \eta s_t^\eta \boldsymbol{\Sigma}_{t+1}^\eta \boldsymbol{g}_t,$$

where

$$\Sigma_{t+1}^\eta = \left(\frac{1}{D^2}\boldsymbol{I} + 2\eta^2 \sum_{s=1}^{t} \boldsymbol{g}_s \boldsymbol{g}_s^\intercal\right)^{-1} \qquad s_t^\eta = 1 + 2\eta \boldsymbol{g}_t^\intercal (\boldsymbol{w}_t^\eta - \boldsymbol{w}_t)$$

Project onto domain:

$$\boldsymbol{w}_{t+1}^\eta = \operatorname*{arg\,min}_{\boldsymbol{u}\in\mathcal{U}} \quad (\boldsymbol{u} - \tilde{\boldsymbol{w}}_{t+1}^\eta)^\intercal (\Sigma_{t+1}^\eta)^{-1} (\boldsymbol{u} - \tilde{\boldsymbol{w}}_{t+1}^\eta)$$

- If master = slave, i.e. $\boldsymbol{w}_t^\eta = \boldsymbol{w}_t$, then is **Online Newton Step**

MetaGrad Slave as Continuous Exponential Weights

Exponential Weights

- Continuous set of experts $\boldsymbol{u} \in \mathbb{R}^d$
- Gaussian prior $P_1 = \mathcal{N}(\mathbf{0}, D^2 \boldsymbol{I})$

$$\mathrm{d}\tilde{P}_{t+1}(\boldsymbol{u}) = \frac{e^{-\ell_t^\eta(\boldsymbol{u})} \mathrm{d}P_t(\boldsymbol{u})}{Z} \qquad \text{(update)}$$

$$P_{t+1} = \min_{P:\ \boldsymbol{\mu}_P \in \mathcal{U}} \mathrm{KL}(P \| \tilde{P}_{t+1}) \qquad \text{(project)}$$

Play the mean of exponential weights:

$$\tilde{P}_t = \mathcal{N}(\tilde{\boldsymbol{w}}_t^\eta, \Sigma_t^\eta)$$
$$P_t = \mathcal{N}(\boldsymbol{w}_t^\eta, \Sigma_t^\eta)$$

- Can understand **Online Newton Step** and many other algorithms this way

MetaGrad Slave Analysis

Standard exponential weights analysis gives **regret bound in space of distributions** for all Q:

$$\mathsf{KL}(Q\|P_1) \geq \sum_{t=1}^{T} -\ln \mathop{\mathbb{E}}_{P_t}[e^{-\ell_t^\eta(\boldsymbol{u})}] - \sum_{t=1}^{T} \mathop{\mathbb{E}}_{Q}[\ell_t^\eta(\boldsymbol{u})]$$

$$\overset{exp-conc}{\geq} \sum_{t=1}^{T} \ell_t^\eta(\boldsymbol{\mu}_{P_t}) - \sum_{t=1}^{T} \mathop{\mathbb{E}}_{Q}[\ell_t^\eta(\boldsymbol{u})]$$

MetaGrad Slave Analysis

Standard exponential weights analysis gives **regret bound in space of distributions** for all Q:

$$\mathrm{KL}(Q\|P_1) \geq \sum_{t=1}^{T} -\ln \mathop{\mathbb{E}}_{P_t}[e^{-\ell_t^\eta(\boldsymbol{u})}] - \sum_{t=1}^{T} \mathop{\mathbb{E}}_{Q}[\ell_t^\eta(\boldsymbol{u})]$$

$$\overset{exp-conc}{\geq} \sum_{t=1}^{T} \ell_t^\eta(\boldsymbol{\mu}_{P_t}) - \sum_{t=1}^{T} \mathop{\mathbb{E}}_{Q}[\ell_t^\eta(\boldsymbol{u})]$$

Specialize to $Q = \mathcal{N}(\boldsymbol{u}^*, D^2\Sigma)$ + **algebra**:

$$\frac{1}{2D^2}\|\boldsymbol{u}^*\|^2 + \frac{1}{2}\left(-\ln\det(\Sigma) + \mathrm{tr}(\Sigma) - d\right)$$

$$\geq \sum_{t=1}^{T} \ell_t^\eta(\boldsymbol{\mu}_{P_t}) - \sum_{t=1}^{T} \ell_t^\eta(\boldsymbol{u}^*) - \sum_{t=1}^{T} \eta^2 D^2\,\mathrm{tr}(\Sigma \boldsymbol{g}_t \boldsymbol{g}_t^\intercal)$$

MetaGrad Slave Analysis

Standard exponential weights analysis gives **regret bound in space of distributions** for all Q:

$$\mathrm{KL}(Q\|P_1) \geq \sum_{t=1}^{T} -\ln \mathop{\mathbb{E}}_{P_t}[e^{-\ell_t^\eta(\boldsymbol{u})}] - \sum_{t=1}^{T} \mathop{\mathbb{E}}_{Q}[\ell_t^\eta(\boldsymbol{u})]$$

$$\overset{exp-conc}{\geq} \sum_{t=1}^{T} \ell_t^\eta(\boldsymbol{\mu}_{P_t}) - \sum_{t=1}^{T} \mathop{\mathbb{E}}_{Q}[\ell_t^\eta(\boldsymbol{u})]$$

Specialize to $Q = \mathcal{N}(\boldsymbol{u}^*, D^2\Sigma)$ + **algebra**:

$$\frac{1}{2D^2}\|\boldsymbol{u}^*\|^2 + \frac{1}{2}\left(-\ln\det(\Sigma) + \mathrm{tr}(\Sigma) - d\right)$$

$$\geq \sum_{t=1}^{T} \ell_t^\eta(\boldsymbol{\mu}_{P_t}) - \sum_{t=1}^{T} \ell_t^\eta(\boldsymbol{u}^*) - \sum_{t=1}^{T} \eta^2 D^2 \,\mathrm{tr}(\Sigma \boldsymbol{g}_t \boldsymbol{g}_t^\intercal)$$

Optimize Σ:

$$R_{\text{slave}}^{\boldsymbol{u}}(\eta) \leq \tfrac{1}{2D^2}\|\boldsymbol{u}^*\| + \tfrac{1}{2}\ln\det\left(\boldsymbol{I} + 2\eta^2 D^2 \sum_{t=1}^{T} \boldsymbol{g}_t \boldsymbol{g}_t^\intercal\right) = O(d \ln T)$$

Outline

Two General Fast Rate Conditions

MetaGrad Algorithm

Exponential Weights Interpretation of Online Learning Algorithms

Many Algorithms as Exponential Weights

Exponentiated Gradient [Kivinen and Warmuth, 1997]

- Continuous set of experts $\boldsymbol{u} \in \mathbb{R}^d$ with loss $\ell_t(\boldsymbol{u}) = \langle \boldsymbol{u}, \boldsymbol{g}_t \rangle$
- Prior P_1 puts point masses on corners of probability simplex

$$\mathrm{d}P_{t+1}(\boldsymbol{u}) = \frac{e^{-\eta \ell_t(\boldsymbol{u})} \mathrm{d}P_t(\boldsymbol{u})}{Z} \qquad \Longrightarrow \qquad \mathop{\mathbb{E}}_{P_{t+1}}[\boldsymbol{u}] = P_{t+1} = \mathrm{EG}$$

Many Algorithms as Exponential Weights

Exponentiated Gradient [Kivinen and Warmuth, 1997]

- Continuous set of experts $\boldsymbol{u} \in \mathbb{R}^d$ with loss $\ell_t(\boldsymbol{u}) = \langle \boldsymbol{u}, \boldsymbol{g}_t \rangle$
- Prior P_1 puts point masses on corners of probability simplex

$$\mathrm{d}P_{t+1}(\boldsymbol{u}) = \frac{e^{-\eta \ell_t(\boldsymbol{u})} \mathrm{d}P_t(\boldsymbol{u})}{Z} \qquad \Longrightarrow \qquad \mathop{\mathbb{E}}_{P_{t+1}}[\boldsymbol{u}] = P_{t+1} = \mathsf{EG}$$

Gradient Descent

Gaussian prior $P_1 = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \quad \Longrightarrow \quad P_{t+1}(\boldsymbol{u}) = \mathcal{N}(-\eta \sum_{t=1}^{T} \boldsymbol{g}_t, \boldsymbol{I})$

Many Algorithms as Exponential Weights

Exponentiated Gradient [Kivinen and Warmuth, 1997]

- Continuous set of experts $\boldsymbol{u} \in \mathbb{R}^d$ with loss $\ell_t(\boldsymbol{u}) = \langle \boldsymbol{u}, \boldsymbol{g}_t \rangle$
- Prior P_1 puts point masses on corners of probability simplex

$$\mathrm{d}P_{t+1}(\boldsymbol{u}) = \frac{e^{-\eta \ell_t(\boldsymbol{u})} \mathrm{d}P_t(\boldsymbol{u})}{Z} \qquad \Longrightarrow \qquad \mathop{\mathbb{E}}_{P_{t+1}}[\boldsymbol{u}] = P_{t+1} = \text{EG}$$

Gradient Descent

$$\text{Gaussian prior } P_1 = \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \quad \Longrightarrow \quad P_{t+1}(\boldsymbol{u}) = \mathcal{N}(-\eta \sum_{t=1}^{T} \boldsymbol{g}_t, \boldsymbol{I})$$

Mirror Descent [Van der Hoeven, 2016]

- Generalize to arbitrary prior P_1
- $\Phi(\theta) = \ln \int e^{\langle \theta, \boldsymbol{u} \rangle} \mathrm{d}P_1(\boldsymbol{u})$:
  CGF for exponential family with carrier P_1

$$\boldsymbol{\mu}_{P_{t+1}} = \nabla \Phi \underbrace{\left(\nabla \Phi^*(\boldsymbol{\mu}_{P_t}) - \eta \boldsymbol{g}_t \right)}_{\text{update in natural parameters}}$$

Dirk

Summary and Last Remarks

MetaGrad

- $\tilde{O}(\sqrt{T})$ regret
- **Fast rates** (often $O(d \ln T)$) for:
  - Adversarial: exp-concave, strongly convex, fixed functions
  - Stochastic: under Bernstein condition (including for hinge loss)

Summary and Last Remarks

MetaGrad

- $\tilde{O}(\sqrt{T})$ regret
- **Fast rates** (often $O(d \ln T)$) for:
  - Adversarial: exp-concave, strongly convex, fixed functions
  - Stochastic: under Bernstein condition (including for hinge loss)

MetaGrad Algorithm

- Master:
  - Pays only $O(\ln \ln T)$ for learning the η that is empirically optimal on the data
  - Almost exponential weights for surrogate loss, but need to **tilt towards larger learning rates**
- Slave:
  - Continuous exponential weights on surrogate loss
  - Matrix updates take $O(d^2)$ work, projections often $O(d^3)$
  - Open problem: add sketching like [Luo et al., 2016]?

Summary and Last Remarks

MetaGrad

- $\tilde{O}(\sqrt{T})$ regret
- **Fast rates** (often $O(d \ln T)$) for:
  - Adversarial: exp-concave, strongly convex, fixed functions
  - Stochastic: under Bernstein condition (including for hinge loss)

MetaGrad Algorithm

- Master:
  - Pays only $O(\ln \ln T)$ for learning the η that is empirically optimal on the data
  - Almost exponential weights for surrogate loss, but need to **tilt towards larger learning rates**
- Slave:
  - Continuous exponential weights on surrogate loss
  - Matrix updates take $O(d^2)$ work, projections often $O(d^3)$
  - Open problem: add sketching like [Luo et al., 2016]?

Cool Aside: View many online algorithms as continuous exponential weights

References

P. L. Bartlett, E. Hazan, and A. Rakhlin. Adaptive online gradient descent. In *NIPS 20*, pages 65–72, 2007.

C.-K. Chiang, T. Yang, C.-J. Le, M. Mahdavi, C.-J. Lu, R. Jin, and S. Zhu. Online optimization with gradual variations. In *Proc. of the 25th Annual Conf. on Learning Theory (COLT)*, pages 6.1–6.20, 2012.

C. B. Do, Q. V. Le, and C.-S. Foo. Proximal regularization for online and batch learning. In *Proc. of the 26th Annual International Conf. on Machine Learning (ICML)*, pages 257–264, 2009.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

E. Hazan and S. Kale. Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine learning*, 80(2-3):165–188, 2010.

E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

W. M. Koolen, P. Grünwald, and T. van Erven. Combining adversarial guarantees and stochastic fast rates in online learning. *NIPS*, 2016.

H. Luo, A. Agarwal, N. Cesa-Bianchi, and J. Langford. Efficient second order online learning by sketching. In *NIPS 29*, 2016.

F. Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. In *NIPS 27*, pages 1116–1124, 2014.

F. Orabona and D. Pál. Coin betting and parameter-free online learning. In *NIPS 29*, 2016.

F. Orabona, K. Crammer, and N. Cesa-Bianchi. A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 99(3):411–435, 2015.

D. van der Hoeven. Is mirror descent a special case of exponential weights? Master's thesis, Leiden University, 2016. Available from `https://www.math.leidenuniv.nl/en/theses/year/2016/`.

T. van Erven and W. M. Koolen. Metagrad: Multiple learning rates in online learning. *NIPS*, 2016.