MetaGrad: Multiple Learning Rates in Online Learning

Tim van Erven   Wouter M. Koolen

Abstract

To get good performance in online convex optimization you need to **select and tune your algorithm** based on **lots of technical stuff**.

Grand goal: single algorithm that works well in all cases.

Multiple Eta Gradient (MetaGrad) algorithm learns **optimal learning rate** from data.

Provable Guarantees:

- Robust to worst-case convex losses
- Adapts to curvature (strong-convex, exp-concave)
- Exploits stochastic data (Bernstein)

Online Convex Optimization Setting

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

Measure **regret** w.r.t. $\boldsymbol{u} \in \mathcal{U}$: $\text{Regret}_T^{\boldsymbol{u}} = \sum_{t=1}^T f_t(\boldsymbol{w}_t) - \sum_{t=1}^T f_t(\boldsymbol{u})$.

Standard Theory

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] handle two cases: **strongly convex + worst-case convex**

MetaGrad Covers Many Cases

Convex f_t	$\sqrt{T \ln \ln T}$
Exp-concave, strongly convex f_t	$d \ln T$
β-Bernstein i.i.d. f_t	$(d \ln T)^{\frac{1}{2-\beta}} \, T^{\frac{1-\beta}{2-\beta}}$

Bernstein distributions with $\beta = 1$ very common:

Absolute loss*	$f_t(u) = \lvert u - X_t \rvert$	$\ln T$
Hinge loss*	$\max\{0, 1 - Y_t \langle \boldsymbol{u}, \boldsymbol{X}_t \rangle\}$	$d \ln T$

Main Theorem

Theorem 1. *MetaGrad's regret is bounded by*

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

where $V_T^{\boldsymbol{u}} = \sum_{t=1}^T \left((\boldsymbol{u} - \boldsymbol{w}_t)^\intercal \boldsymbol{g}_t\right)^2$.

Fast Rates: Directional Derivative Condition

Theorem 2. *If there exist $a, b > 0$ such that all f_t satisfy*

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

then $\text{Regret}_T^{\boldsymbol{u}} \le O(d \ln T)$.

- Satisfied by **exp-concave** and **strongly convex** functions [Hazan, Agarwal, and Kale, 2007] with $a = 1$.
- Satisfied for any **fixed convex** function $f_t = f$ with minimizer $\boldsymbol{u}$, even without any curvature, with $a = 2$.

Fast Rates: Stochastic Bernstein Condition

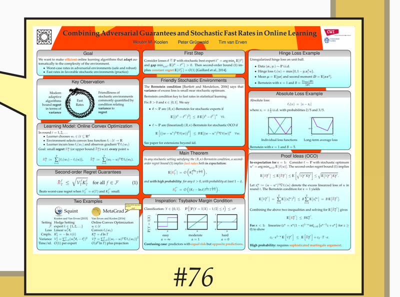
Consider $f_t \overset{\text{iid}}{\sim} \mathbb{P}$ with stochastic optimum $\boldsymbol{u}^* = \arg\min_{\boldsymbol{u} \in \mathcal{U}} \mathbb{E}_f[f(\boldsymbol{u})]$ satisfying the (linearized) **(B, β)-Bernstein condition**

$$\mathbb{E}\left[\left((\boldsymbol{w} - \boldsymbol{u}^*)^\intercal \nabla f(\boldsymbol{w})\right)^2\right] \;\le\; B\, \mathbb{E}\left[(\boldsymbol{w} - \boldsymbol{u}^*)^\intercal \nabla f(\boldsymbol{w})\right]^\beta \quad \forall \boldsymbol{w} \in \mathcal{U}.$$

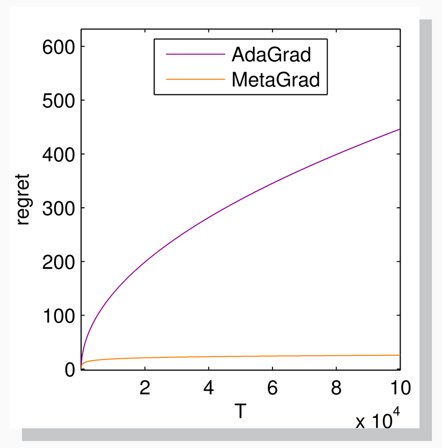
Example: Hinge loss (unit ball): $\beta = 1$, $B = \frac{2\lambda_{\max}(\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\intercal])}{\lVert \mathbb{E}[Y\boldsymbol{X}] \rVert}$

Theorem 3 (Koolen, Grünwald, Van Erven, 2016).

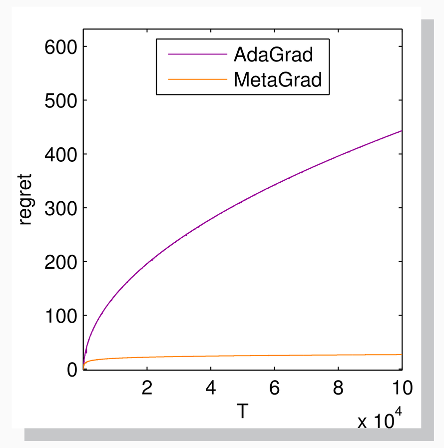
$$\mathbb{E}[\text{Regret}_T^{\boldsymbol{u}^*}] = O\left((B d \ln T)^{\frac{1}{2-\beta}} \, T^{\frac{1-\beta}{2-\beta}} \right)$$



Experiments (Proof-of-Concept)



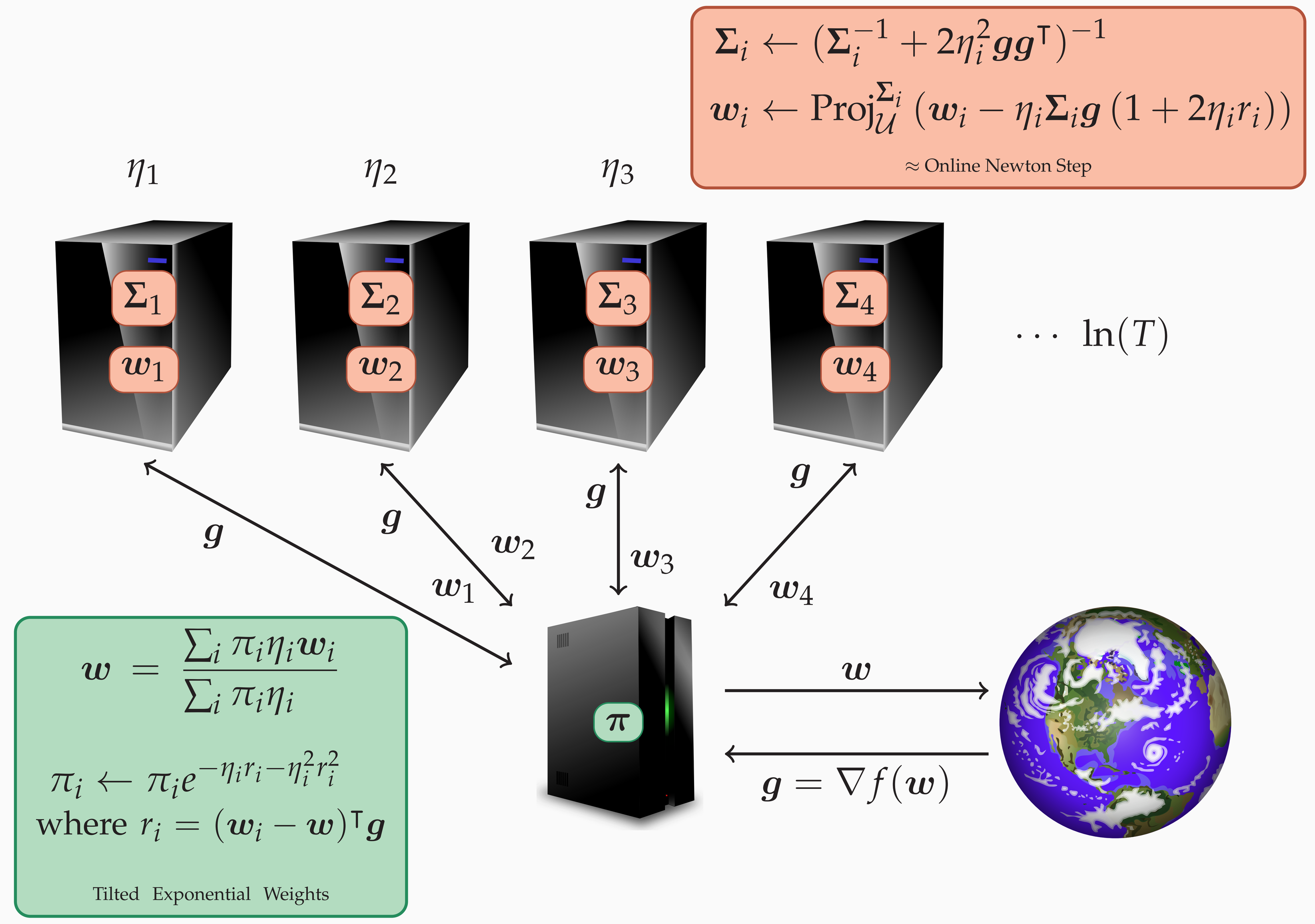
Offline: fixed $f_t(u) = \lvert u - 1/4 \rvert$



Stochastic Online: $f_t(u) = \lvert u - X_t \rvert$ where $X_t = \pm 1/2$ i.i.d. w.p. 2/5 and 3/5.

- MetaGrad: $O(\ln T)$ regret, AdaGrad: $O(\sqrt{T})$, match bounds
- Functions neither strongly convex nor smooth

MetaGrad Algorithm



Proof Ideas

Analysis based on second-order **surrogate loss**. For each η:

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

Since surrogate is **exp-concave** for each fixed η, we can use **online quasi-Newton method** like Online Newton Step [Hazan et al., 2007] to get predictions $\boldsymbol{w}_t^\eta$ that achieve logarithmic regret:

$$\sum_{t=1}^T \ell_t^\eta(\boldsymbol{w}_t^\eta) - \sum_{t=1}^T \ell_t^\eta(\boldsymbol{u}) \;\le\; O(d \ln T) \qquad \forall \boldsymbol{u} \in \mathcal{U}$$

To learn the **best** η we combine the predictions $\boldsymbol{w}_t^\eta$ for multiple η into a single master prediction $\boldsymbol{w}_t$ using an **experts algorithm for combining multiple learning rates** similar to Squint [Koolen and Van Erven, 2015], to get:

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

Difficulty: Master has to perform well under multiple loss functions simultaneously. No standard experts algorithm works!

Together: $-\sum_{t=1}^T \ell_t^\eta(\boldsymbol{u}) \le O(d \ln T)$ for each η and $\boldsymbol{u}$, resulting in

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