MetaGrad: Multiple Learning Rates in Online Learning

Tim van Erven

Universiteit Leiden

Joint work with: Wouter Koolen, Peter Grünwald

Optimization and Learning Workshop
Toulouse, September 11, 2018

Example: Sequential Prediction for Football Games

Precursor to modern football in China,
Han Dynasty (206 BC – 220 AD)

- Before every match t in the English Premier League, my PhD student Dirk van der Hoeven wants to predict the goal difference Y_t
- Given feature vector $\boldsymbol{X}_t \in \mathbb{R}^d$, he may predict $\hat{Y}_t = \boldsymbol{w}_t^\intercal \boldsymbol{X}_t$ with a linear model
- After the match: observe Y_t
- Measure loss by $\ell_t(\boldsymbol{w}_t) = (Y_t - \hat{Y}_t)^2$ and improve parameter estimates: $\boldsymbol{w}_t \to \boldsymbol{w}_{t+1}$

Example: Sequential Prediction for Football Games

Precursor to modern football in China,
Han Dynasty (206 BC – 220 AD)

- Before every match t in the English Premier League, my PhD student Dirk van der Hoeven wants to predict the goal difference Y_t
- Given feature vector $\boldsymbol{X}_t \in \mathbb{R}^d$, he may predict $\hat{Y}_t = \boldsymbol{w}_t^\intercal \boldsymbol{X}_t$ with a linear model
- After the match: observe Y_t
- Measure loss by $\ell_t(\boldsymbol{w}_t) = (Y_t - \hat{Y}_t)^2$ and improve parameter estimates: $\boldsymbol{w}_t \to \boldsymbol{w}_{t+1}$

Goal: Predict almost as well as the best possible parameters $\boldsymbol{u}$:

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

Online Convex Optimization

```
for $t = 1, 2, \ldots, T$ do
    Learner estimates $\boldsymbol{w}_t$ from convex $\mathcal{U} \subset \mathbb{R}^d$
    Nature reveals convex loss function $\ell_t : \mathcal{U} \to \mathbb{R}$
    Learner incurs loss $\ell_t(\boldsymbol{w}_t)$
end for
```

Online Convex Optimization

```
for $t = 1, 2, \ldots, T$ do
    Learner estimates $\boldsymbol{w}_t$ from convex $\mathcal{U} \subset \mathbb{R}^d$
    Nature reveals convex loss function $\ell_t : \mathcal{U} \to \mathbb{R}$
    Learner incurs loss $\ell_t(\boldsymbol{w}_t)$
end for
```

Viewed as a **zero-sum game** against Nature:

$$V = \min_{\boldsymbol{w}_1} \max_{\ell_1} \min_{\boldsymbol{w}_2} \max_{\ell_2} \cdots \min_{\boldsymbol{w}_T} \max_{\ell_T} \max_{\boldsymbol{u} \in \mathcal{U}} \underbrace{\sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u})}_{\text{Regret}_T^{\boldsymbol{u}}}$$

Online Convex Optimization

```
1: for $t = 1, 2, \ldots, T$ do
2:     Learner estimates $\boldsymbol{w}_t$ from convex $\mathcal{U} \subset \mathbb{R}^d$
3:     Nature reveals convex loss function $\ell_t : \mathcal{U} \to \mathbb{R}$
4:     Learner incurs loss $\ell_t(\boldsymbol{w}_t)$
5: end for
```

Viewed as a **zero-sum game** against Nature:

$$V = \min_{\boldsymbol{w}_1} \max_{\ell_1} \min_{\boldsymbol{w}_2} \max_{\ell_2} \cdots \min_{\boldsymbol{w}_T} \max_{\ell_T} \max_{\boldsymbol{u} \in \mathcal{U}} \underbrace{\sum_{t=1}^{T} \ell_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} \ell_t(\boldsymbol{u})}_{\text{Regret}_T^{\boldsymbol{u}}}$$

Methods: Efficient computations using only gradient $\boldsymbol{g}_t = \nabla \ell_t(\boldsymbol{w}_t)$

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \boldsymbol{g}_t \qquad \text{(online gradient descent)}$$

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \Sigma_{t+1} \boldsymbol{g}_t \qquad \text{(online Newton Step)}$$

where $\Sigma_{t+1} = (\epsilon I + 2\eta^2 \sum_{s=1}^{t} \boldsymbol{g}_s \boldsymbol{g}_s^\intercal)^{-1}$.

The Standard Picture

Minimax rates based on curvature (bounded domain and gradients) [Hazan, 2016]:

Convex ℓ_t	$\sqrt{T}$	OGD with $\eta_t \propto \frac{1}{\sqrt{t}}$
Strongly convex ℓ_t	$\ln T$	OGD with $\eta_t \propto \frac{1}{t}$
Exp-concave ℓ_t	$d \ln T$	ONS with $\eta \propto 1$

- **Strongly convex:** second derivative at least $\alpha > 0$, implies exp-concave
- **Exp-concave:** $e^{-\alpha \, \ell_t}$ concave
  Satisfied by log loss, logistic loss, squared loss, but not hinge loss

The Standard Picture

Minimax rates based on curvature (bounded domain and gradients) [Hazan, 2016]:

Convex ℓ_t	$\sqrt{T}$	OGD with $\eta_t \propto \frac{1}{\sqrt{t}}$
Strongly convex ℓ_t	$\ln T$	OGD with $\eta_t \propto \frac{1}{t}$
Exp-concave ℓ_t	$d \ln T$	ONS with $\eta \propto 1$

Limitations:

- Different method in each case. (Requires sophisticated users.)
- Theoretical tuning of η_t **very conservative**
- What if curvature varies between rounds?
- In many applications data are **stochastic** (i.i.d.) Should be easier than worst case. . .

The Standard Picture

Minimax rates based on curvature (bounded domain and gradients) [Hazan, 2016]:

Convex ℓ_t	$\sqrt{T}$	OGD with $\eta_t \propto \frac{1}{\sqrt{t}}$
Strongly convex ℓ_t	$\ln T$	OGD with $\eta_t \propto \frac{1}{t}$
Exp-concave ℓ_t	$d \ln T$	ONS with $\eta \propto 1$

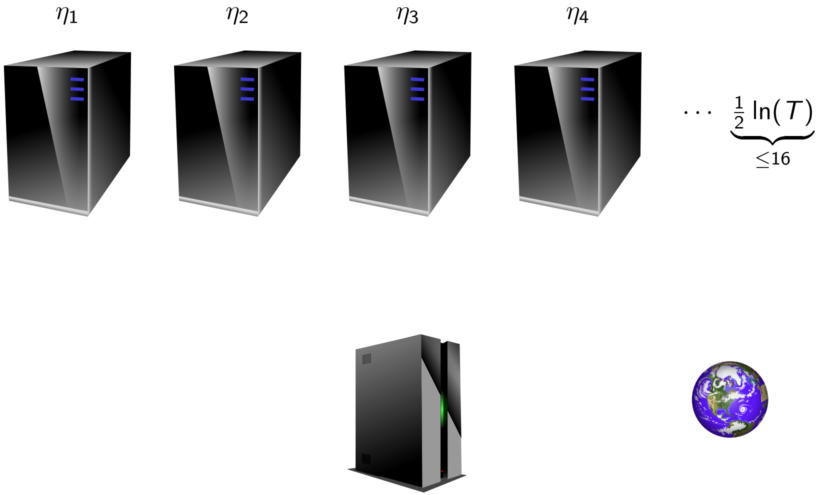
Limitations:

- Different method in each case. (Requires sophisticated users.)
- Theoretical tuning of η_t **very conservative**
- What if curvature varies between rounds?
- In many applications data are **stochastic** (i.i.d.) Should be easier than worst case. . .

Need Adaptive Methods!

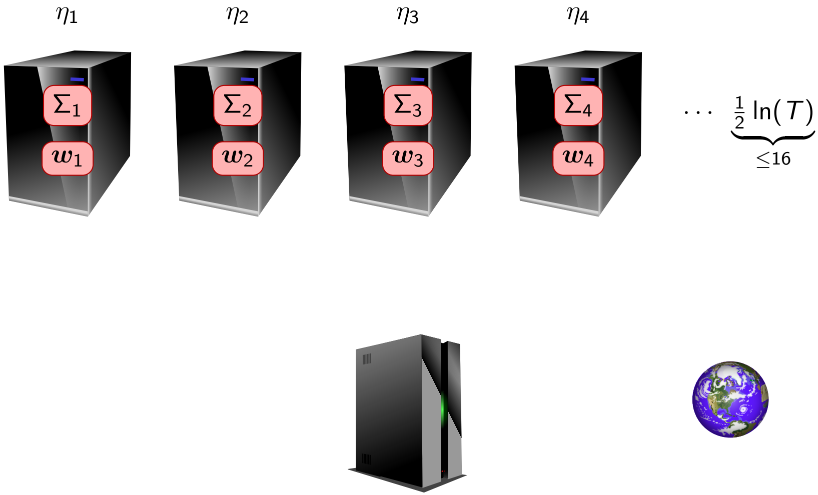
- Difficulty: All existing methods learn η at too slow rate [HP2005] so **overhead of learning best η ruins potential benefits**

MetaGrad: Multiple Eta Gradient Algorithm

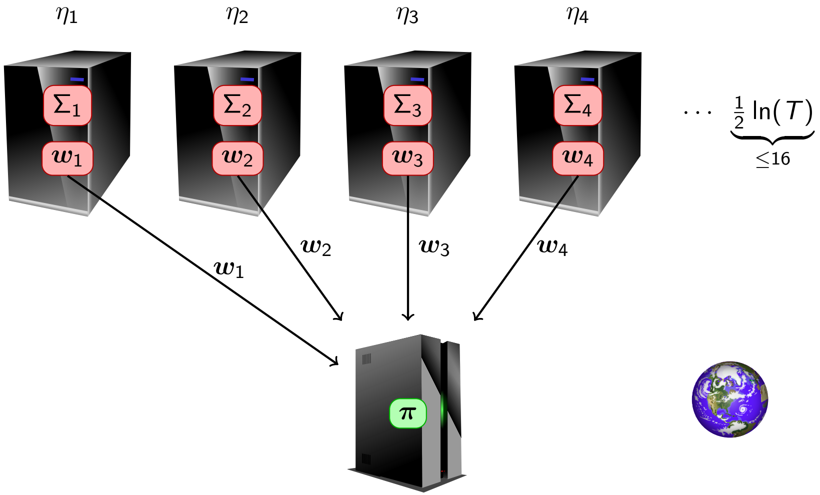


$\cdots \quad \underbrace{\tfrac{1}{2}\ln(T)}_{\leq 16}$

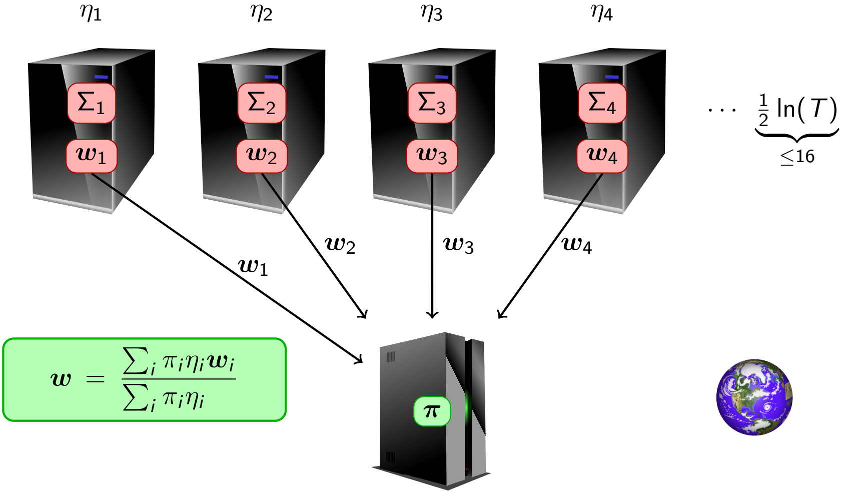
MetaGrad: Multiple Eta Gradient Algorithm



MetaGrad: Multiple Eta Gradient Algorithm

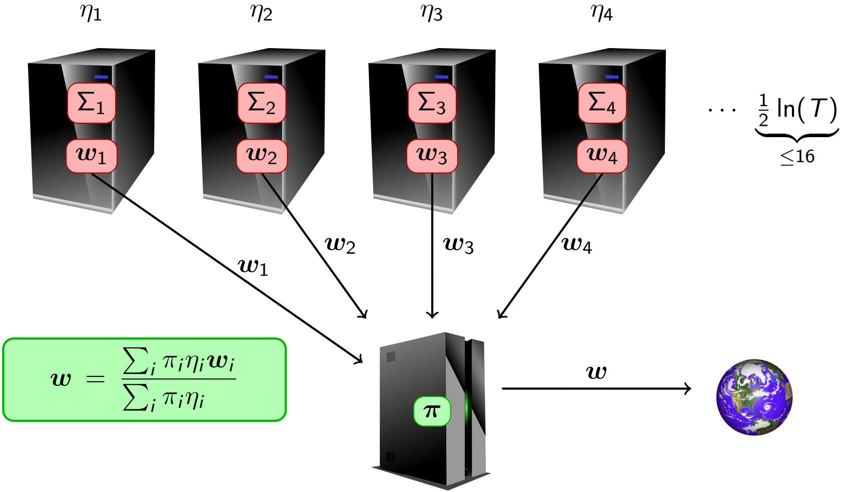


MetaGrad: Multiple Eta Gradient Algorithm



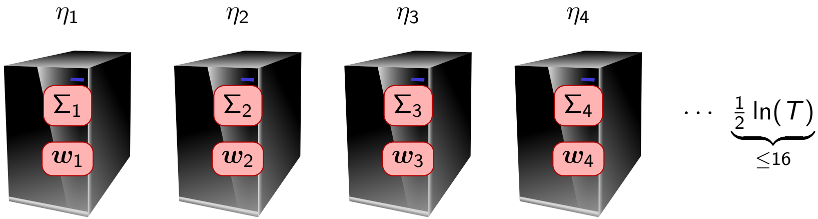
$$\boldsymbol{w} = \frac{\sum_i \pi_i \eta_i \boldsymbol{w}_i}{\sum_i \pi_i \eta_i}$$

MetaGrad: Multiple Eta Gradient Algorithm

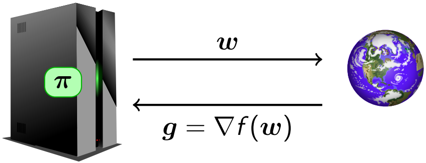


$$\boldsymbol{w} = \frac{\sum_i \pi_i \eta_i \boldsymbol{w}_i}{\sum_i \pi_i \eta_i}$$

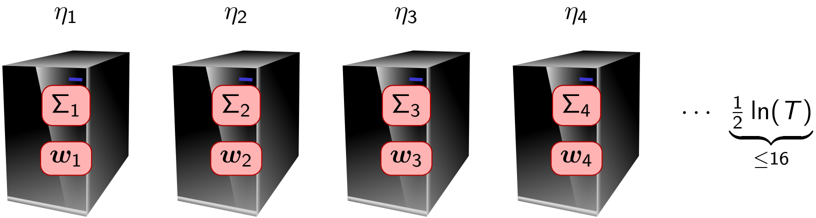
MetaGrad: Multiple Eta Gradient Algorithm



$$\boxed{\boldsymbol{w} = \frac{\sum_i \pi_i \eta_i \boldsymbol{w}_i}{\sum_i \pi_i \eta_i}}$$



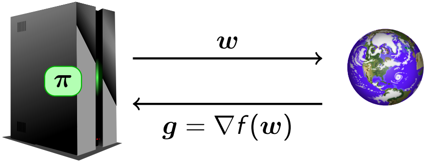
MetaGrad: Multiple Eta Gradient Algorithm



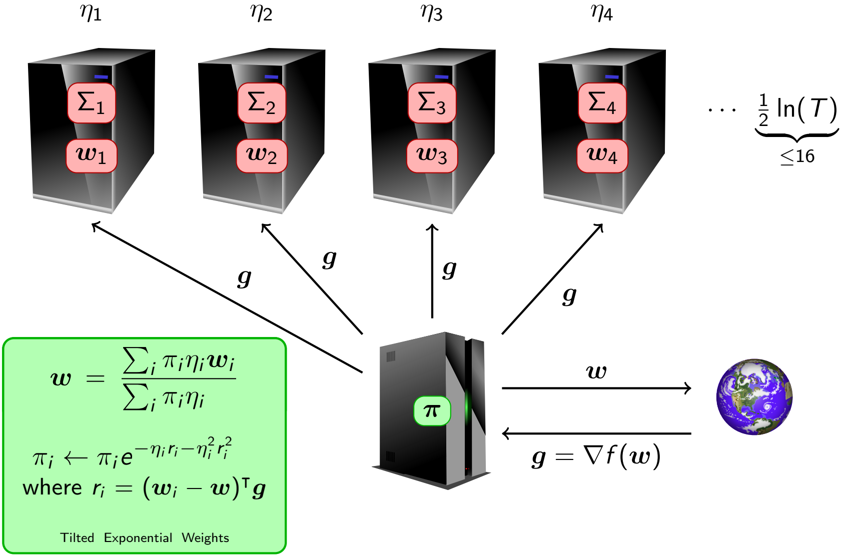
$$\boldsymbol{w} = \frac{\sum_i \pi_i \eta_i \boldsymbol{w}_i}{\sum_i \pi_i \eta_i}$$

$\pi_i \leftarrow \pi_i e^{-\eta_i r_i - \eta_i^2 r_i^2}$
where $r_i = (\boldsymbol{w}_i - \boldsymbol{w})^\intercal \boldsymbol{g}$

Tilted Exponential Weights



MetaGrad: Multiple Eta Gradient Algorithm



$$\boldsymbol{w} = \frac{\sum_i \pi_i \eta_i \boldsymbol{w}_i}{\sum_i \pi_i \eta_i}$$

$$\pi_i \leftarrow \pi_i e^{-\eta_i r_i - \eta_i^2 r_i^2}$$

where $r_i = (\boldsymbol{w}_i - \boldsymbol{w})^\intercal \boldsymbol{g}$

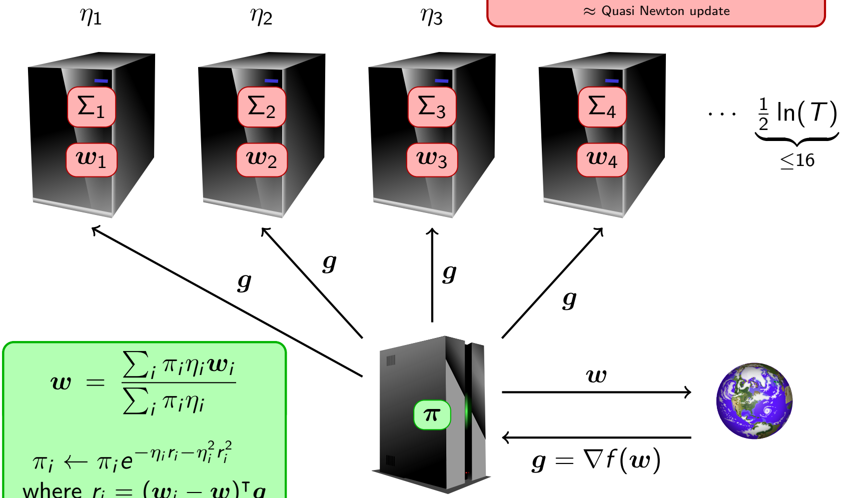
Tilted Exponential Weights

MetaGrad: Multiple Eta G

$$\Sigma_i \leftarrow (\Sigma_i^{-1} + 2\eta_i^2 \boldsymbol{g}\boldsymbol{g}^\intercal)^{-1}$$

$$\boldsymbol{w}_i \leftarrow \boldsymbol{w}_i - \eta_i \Sigma_i \boldsymbol{g}(1 + 2\eta_i r_i)$$

≈ Quasi Newton update



$$\boldsymbol{w} = \frac{\sum_i \pi_i \eta_i \boldsymbol{w}_i}{\sum_i \pi_i \eta_i}$$

$$\pi_i \leftarrow \pi_i e^{-\eta_i r_i - \eta_i^2 r_i^2}$$

where $r_i = (\boldsymbol{w}_i - \boldsymbol{w})^\intercal \boldsymbol{g}$

Tilted Exponential Weights

MetaGrad: Provable Adaptive Fast Rates

Theorem (Van Erven, Koolen, 2016)

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

$$\text{Regret}_T^{\boldsymbol{u}} \leq \sum_{t=1}^{T} (\boldsymbol{w}_t - \boldsymbol{u})^\intercal \boldsymbol{g}_t \preccurlyeq \begin{cases} \sqrt{T \ln \ln T} \\ \\ \sqrt{V_T^{\boldsymbol{u}}\, d \ln T} + d \ln T \end{cases}$$

where

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

- By convexity, $\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) \leq (\boldsymbol{w}_t - \boldsymbol{u})^\intercal \boldsymbol{g}_t$.
- Optimal learning rate η depends on $V_T^{\boldsymbol{u}}$, but $\boldsymbol{u}$ unknown!
  Crucial to learn best learning rate from data!

MetaGrad: Provable Adaptive Fast Rates

Theorem (Van Erven, Koolen, 2016)

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

$$\text{Regret}_T^{\boldsymbol{u}} \leq \sum_{t=1}^{T} (\boldsymbol{w}_t - \boldsymbol{u})^\intercal \boldsymbol{g}_t \preccurlyeq \begin{cases} \sqrt{T \ln \ln T} \\ \\ \sqrt{V_T^{\boldsymbol{u}} \, d \ln T} + d \ln T \end{cases}$$

where

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

- By convexity, $\ell_t(\boldsymbol{w}_t) - \ell_t(\boldsymbol{u}) \leq (\boldsymbol{w}_t - \boldsymbol{u})^\intercal \boldsymbol{g}_t$.
- Optimal learning rate η depends on $V_T^{\boldsymbol{u}}$, but $\boldsymbol{u}$ unknown!
  Crucial to learn best learning rate from data!

Consequences

1. Non-stochastic adaptation:

Convex ℓ_t	$\sqrt{T \ln \ln T}$
Exp-concave ℓ_t	$d \ln T$
Fixed convex $\ell_t = \ell$	$d \ln T$

Consequences

1. Non-stochastic adaptation:

Convex ℓ_t	$\sqrt{T \ln \ln T}$
Exp-concave ℓ_t	$d \ln T$
Fixed convex $\ell_t = \ell$	$d \ln T$

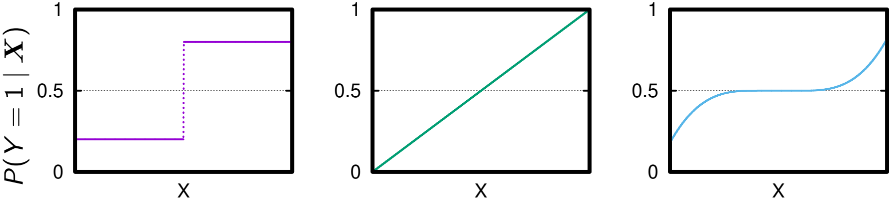
2. Stochastic without curvature

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

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

*Conditions apply

Related Work: Adaptivity to Stochastic Data in Batch Classification [Tsybakov, 2004]

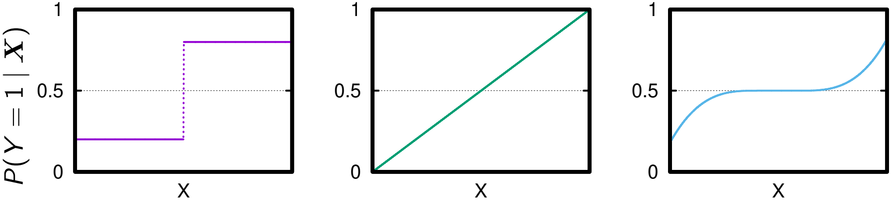


easy
$\beta = 1$

moderate
$\beta = \frac{1}{2}$

hard
$\beta = 0$

Related Work: Adaptivity to Stochastic Data in Batch Classification [Tsybakov, 2004]



easy
$\beta = 1$

moderate
$\beta = \frac{1}{2}$

hard
$\beta = 0$

Definition ((B, β)-Bernstein Condition)

Losses are i.i.d. and

$$\mathbb{E}\left(\ell(\boldsymbol{w}) - \ell(\boldsymbol{u}^*)\right)^2 \leq B\Big(\mathbb{E}\left[\ell(\boldsymbol{w}) - \ell(\boldsymbol{u}^*)\right]\Big)^{\beta} \qquad \text{for all } \boldsymbol{w},$$

where $\boldsymbol{u}^* = \arg\min_{\boldsymbol{u}} \mathbb{E}[\ell(\boldsymbol{u})]$ minimizes the expected loss.

Bernstein Condition for Online Learning

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

Standard Bernstein condition:

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

Bernstein Condition for Online Learning

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

Standard Bernstein condition:

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

Replace by **weaker linearized version:**

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

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

Bernstein Condition for Online Learning

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

Standard Bernstein condition:

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

Replace by **weaker linearized version:**

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

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

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

Bernstein Condition for Online Learning

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

Standard Bernstein condition:

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

Replace by **weaker linearized version:**

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

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

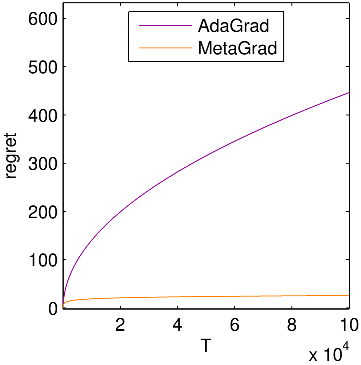
Hinge loss (domain, gradients bounded by 1): $\beta = 1$, $B = \frac{2\lambda_{\max}(\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\intercal])}{\|\mathbb{E}[Y\boldsymbol{X}]\|}$

Theorem (Koolen, Grünwald, Van Erven, 2016)

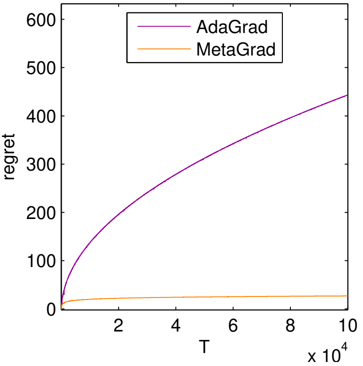
$$\mathbb{E}[\text{Regret}_T^{\boldsymbol{u}^*}] \preccurlyeq (Bd \ln T)^{1/(2-\beta)}\, T^{(1-\beta)/(2-\beta)}$$

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

MetaGrad Simulation Experiments



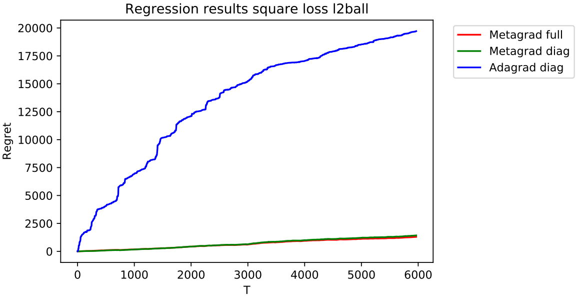
Offline: $\ell_t(u) = |u - 1/4|$



Stochastic Online: $\ell_t(u) = |u - X_t|$
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

MetaGrad Football Experiments



Dirk van der Hoeven
(my PhD student)

Raphaël Deswarte
(visiting PhD student)

- Predict difference in goals in 6000 football games in English Premier League (Aug 2000–May 2017).
- Square loss on Euclidean ball
- 37 features: running average of goals, shots on goal, shots over $m = 1, \dots, 10$ previous games; multiple ELO-like models; intercept.

Analysis

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

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

Analysis

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

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

Analysis

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

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

Analysis

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

Analysis

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

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

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

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

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

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

Analysis

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

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

Analysis

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:

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

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

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

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:

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

Summary

MetaGrad:

- Consider **multiple learning rates** η simultaneously
- Learn η from the data, at very fast rate (pay only $\ln\ln T$)
- New adaptive variance bound

Variance bound implies fast rates in:

- all known cases: exp-concave, strong convex
- new cases with stochastic data, characterized by online version of Bernstein condition

References

- T. van Erven and W. M. Koolen. **Metagrad: Multiple learning rates in online learning.** In Advances in Neural Information Processing Systems 29 (NIPS), pages 3666–3674, 2016.
- W. M. Koolen, P. Grünwald, and T. van Erven. **Combining adversarial guarantees and stochastic fast rates in online learning.** In Advances in Neural Information Processing Systems 29 (NIPS), pages 4457–4465, 2016.

P. L. Bartlett, E. Hazan, and A. Rakhlin. Adaptive online gradient descent. In *Advances in Neural Information Processing Systems 20 (NIPS)*, pages 65–72, 2007.

C. B. Do, Q. V. Le, and C.-S. Foo. Proximal regularization for online and batch learning. In *Proceedings of the 26th Annual International Conference 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. Introduction to online optimization. Draft, April 10, 2016, available from `ocobook.cs.princeton.edu`, 2016.

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

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.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.