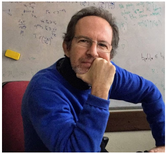
Multitask Online Mirror Descent

P. Laforgue, Università degli Studi di Milano, Milan, Italy



N. Cesa-Bianchi
(UniMi)

A. Paudice
(UniMi & IIT)

M. Pontil
(IIT & UCL)

Online Multitask Learning: Motivations

- Datastreams are ubiquitous: markets, sensors, user interactions
- Many problems are **multitask**: stock predictions, federated learning for mobile users, for smart homes, weather forecasting
- **Is it possible to improve when we face similar tasks?**

Partial **yes** in [Cavallanti et al. 2010] (specific algorithm, loss, geometry)

Online Convex Optimization (single task)

At each time step $t = 1, \dots, T$, the learner:

1. makes a prediction $x_t \in V \subset \mathbb{R}^d$,
2. receives a convex loss function $\ell_t \colon V \to \mathbb{R}$,
3. pays $\ell_t(x_t)$, and uses the knowledge of ℓ_t for the next predictions.

Given a sequence of losses ℓ_t (possibly arbitrary), the goal is to minimize the **regret**, defined as:

$$R_T = \sum_{t=1}^{T} \ell_t(x_t) - \underbrace{\inf_{u \in V} \sum_{t=1}^{T} \ell_t(u)}_{\text{best model in hindsight}}$$

Online Mirror Descent (1/2)

Given $\psi\colon \mathbb{R}^d \to \mathbb{R}$, λ-strongly convex w.r.t. norm $\|\cdot\|$ on V, the OMD update writes:

$$x_{t+1} = \underset{x \in V}{\operatorname{argmin}} \ \langle \eta_t g_t, x \rangle + B_\psi(x, x_t) \tag{1}$$

where $g_t \in \partial \ell_t(x_t)$, and $B_\psi(x,y) = \psi(x) - \psi(y) - \langle \nabla \psi(y), x - y \rangle$.

For $\eta_t \coloneqq \eta$ and any $x_1 \in V$, it can be shown that the sequence of iterates produced by (1) satisfies:

$$\forall u \in V, \qquad R_T(u) \leq \frac{B_\psi(u, x_1)}{\eta} + \frac{\eta}{2\lambda} \sum_{t=1}^{T} \|g_t\|_\star^2$$

Online Mirror Descent (2/2)

$$\forall u \in V, \qquad R_T(u) \le \frac{B_\psi(u, x_1)}{\eta} + \frac{\eta}{2\lambda} \sum_{t=1}^{T} \|g_t\|_\star^2$$

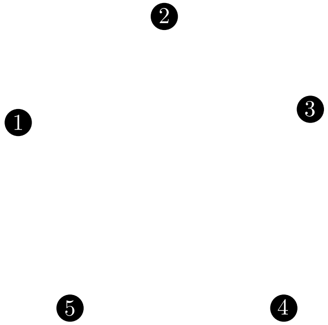
Two famous instances of OMD are Online Gradient Descent (OGD) and Exponentiated Gradient (EG).

	OGD	EG
$\psi(x)$	$\frac{1}{2}\|x\|_2^2$	$\sum_{j=1}^{d} x_j \ln x_j$
$\lambda, \|\cdot\|, \|\cdot\|_\star$	$1, \|\cdot\|_2, \|\cdot\|_2$	$1, \|\cdot\|_1, \|\cdot\|_\infty$
$B_\psi(x, y)$	$\frac{1}{2}\|x - y\|_2^2$	$\sum_{j=1}^{d} x_j \ln\left(\frac{x_j}{y_j}\right)$
R_T on the simplex with $\|g_t\|_\infty \le 1$	$\mathcal{O}(\sqrt{Td})$	$\mathcal{O}(\sqrt{T \ln d})$

Online Multitask Learning: A Multiagent Formalism

N agents, each trying to solve its own task. At time step t, agent i_t is active (arbitrary chosen). Our goal is to minimize the **multitask regret**:

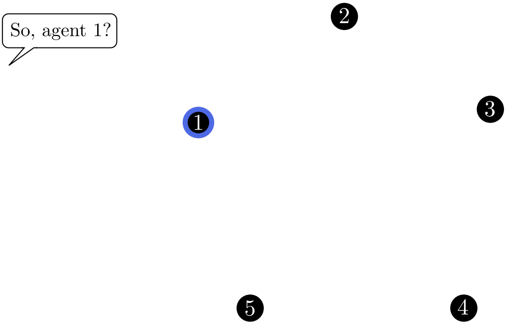
$$\boldsymbol{R}_T = \sum_{i=1}^{N} \left(\sum_{t\,:\; i_t = i} \ell_t(x_t) - \inf_{u \in V} \sum_{t\,:\; i_t = i} \ell_t(u) \right)$$



Online Multitask Learning: A Multiagent Formalism

N agents, each trying to solve its own task. At time step t, agent i_t is active (arbitrary chosen). Our goal is to minimize the **multitask regret**:

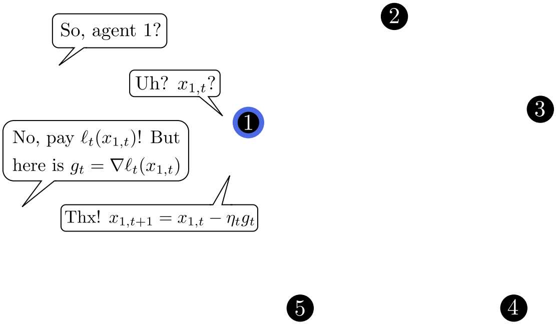
$$\boldsymbol{R}_T = \sum_{i=1}^{N} \left(\sum_{t\colon\, i_t = i} \ell_t(x_t) - \inf_{u \in V} \sum_{t\colon\, i_t = i} \ell_t(u) \right)$$



Online Multitask Learning: A Multiagent Formalism

N agents, each trying to solve its own task. At time step t, agent i_t is active (arbitrary chosen). Our goal is to minimize the **multitask regret**:

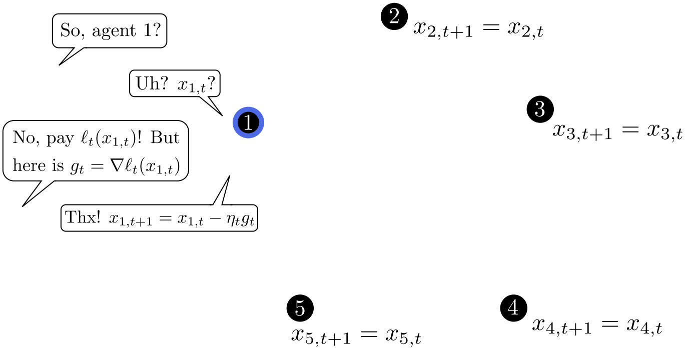
$$\boldsymbol{R}_T = \sum_{i=1}^{N} \left(\sum_{t\,:\ i_t = i} \ell_t(x_t) - \inf_{u \in V} \sum_{t\,:\ i_t = i} \ell_t(u) \right)$$



Online Multitask Learning: A Multiagent Formalism

N agents, each trying to solve its own task. At time step t, agent i_t is active (arbitrary chosen). Our goal is to minimize the **multitask regret**:

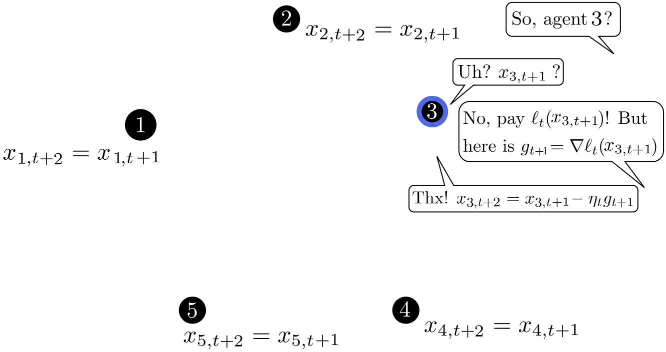
$$\boldsymbol{R}_T = \sum_{i=1}^{N} \left(\sum_{t\,:\; i_t = i} \ell_t(x_t) - \inf_{u \in V} \sum_{t\,:\; i_t = i} \ell_t(u) \right)$$



Online Multitask Learning: A Multiagent Formalism

N agents, each trying to solve its own task. At time step t, agent i_t is active (arbitrary chosen). Our goal is to minimize the **multitask regret**:

$$\boldsymbol{R}_T = \sum_{i=1}^{N} \left(\sum_{t:\ i_t=i} \ell_t(x_t) - \inf_{u \in V} \sum_{t:\ i_t=i} \ell_t(u) \right)$$



Naive Approach: Independent OMDs

If individual OMD has regret bounded by $C\sqrt{T}$, by Jensen's inequality:

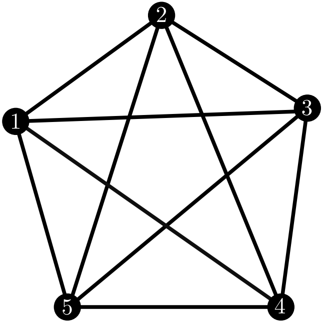
$$\boldsymbol{R}_T \leq \sum_{i=1}^{N} C\sqrt{T_i} \leq C\sqrt{NT}\,.$$

Is it possible to improve with respect to the $\sqrt{N}$ dependence? Yes

How? Under which condition on the tasks? on ψ?

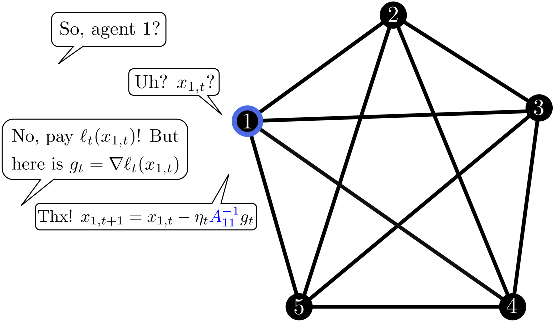
Our approach: Multitask OMD

How? **By sharing gradients between agents**



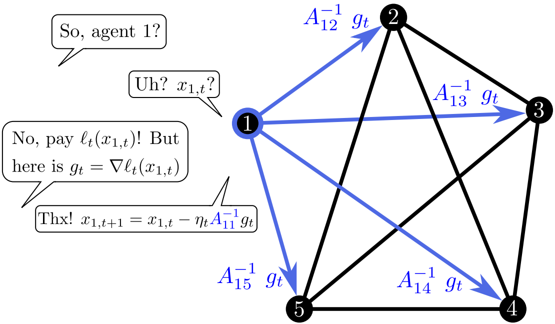
Our approach: Multitask OMD

How? By sharing gradients between agents



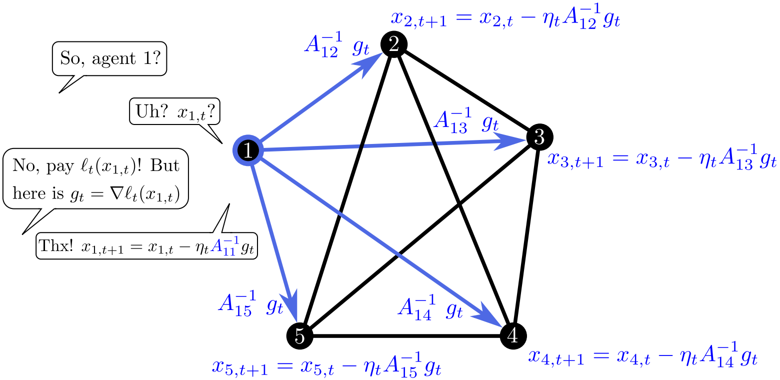
Multitask OMD: our approach

How? **By sharing gradients between agents**



Multitask OMD: our approach

How? By sharing gradients between agents



MT-OMD: Analysis

Let $A \in \mathbb{R}^{N\times N}$, $\boldsymbol{A} = A \otimes I_d \in \mathbb{R}^{Nd\times Nd}$. For a regularizer $\psi\colon \mathbb{R}^d \to \mathbb{R}$, let

$$\boldsymbol{\psi}\colon \boldsymbol{u} \in \mathbb{R}^{Nd} \mapsto \sum_{i=1}^{N} \psi(\boldsymbol{u}^{(i)}), \qquad \tilde{\boldsymbol{\psi}}\colon \boldsymbol{u} \in \mathbb{R}^{Nd} \mapsto \boldsymbol{\psi}(\boldsymbol{A}^{1/2}\boldsymbol{u})$$

We have $B_{\tilde{\boldsymbol{\psi}}}(\boldsymbol{x}, \boldsymbol{y}) = B_{\boldsymbol{\psi}}(\boldsymbol{A}^{1/2}\boldsymbol{x}, \boldsymbol{A}^{1/2}\boldsymbol{y})$, so the MT-OMD update writes:

$$\boldsymbol{x}_{t+1} = \operatorname*{argmin}_{\boldsymbol{x}\in\boldsymbol{V}} \langle \eta_t \bar{g}_t, \boldsymbol{x}\rangle + B_{\boldsymbol{\psi}}(\boldsymbol{A}^{1/2}\boldsymbol{x}, \boldsymbol{A}^{1/2}\boldsymbol{x}_t)$$

$$= \boldsymbol{A}^{-1/2} \operatorname*{argmin}_{\boldsymbol{y}\in\boldsymbol{A}^{1/2}(\boldsymbol{V})} \langle \eta_t \boldsymbol{A}^{-1/2}\bar{g}_t, \boldsymbol{y}\rangle + B_{\boldsymbol{\psi}}(\boldsymbol{y}, \boldsymbol{y}_t)$$

We have shown that:

$$\forall \boldsymbol{u} \in \mathbb{R}^{Nd}, \quad \boldsymbol{R}_T(\boldsymbol{u}) \le \frac{B_{\boldsymbol{\psi}}(\boldsymbol{A}^{1/2}\boldsymbol{u}, \boldsymbol{A}^{1/2}\boldsymbol{x}_1)}{\eta} + \eta \max_{i\le N} A_{ii}^{-1} \sum_{t=1}^{T} \frac{\|g_t\|_*^2}{2\lambda}$$

MT-OMD: Analysis

Let $A \in \mathbb{R}^{N \times N}$, $\boldsymbol{A} = A \otimes I_d \in \mathbb{R}^{Nd \times Nd}$. For a regularizer $\psi\colon \mathbb{R}^d \to \mathbb{R}$, let

$$\boldsymbol{\psi}\colon \boldsymbol{u} \in \mathbb{R}^{Nd} \mapsto \sum_{i=1}^{N} \psi(\boldsymbol{u}^{(i)}), \qquad \tilde{\boldsymbol{\psi}}\colon \boldsymbol{u} \in \mathbb{R}^{Nd} \mapsto \boldsymbol{\psi}(\boldsymbol{A}^{1/2}\boldsymbol{u})$$

We have $B_{\tilde{\boldsymbol{\psi}}}(\boldsymbol{x}, \boldsymbol{y}) = B_{\boldsymbol{\psi}}(\boldsymbol{A}^{1/2}\boldsymbol{x}, \boldsymbol{A}^{1/2}\boldsymbol{y})$, so the MT-OMD update writes:

$$\begin{aligned}\boldsymbol{x}_{t+1} &= \underset{\boldsymbol{x} \in \boldsymbol{V}}{\operatorname{argmin}} \ \langle \eta_t \bar{\boldsymbol{g}}_t, \boldsymbol{x} \rangle + B_{\boldsymbol{\psi}}(\boldsymbol{A}^{1/2}\boldsymbol{x}, \boldsymbol{A}^{1/2}\boldsymbol{x}_t) \\ &= \boldsymbol{A}^{-1/2} \underset{\boldsymbol{y} \in \boldsymbol{A}^{1/2}(\boldsymbol{V})}{\operatorname{argmin}} \ \langle \eta_t \boldsymbol{A}^{-1/2}\bar{\boldsymbol{g}}_t, \boldsymbol{y} \rangle + B_{\boldsymbol{\psi}}(\boldsymbol{y}, \boldsymbol{y}_t)\end{aligned}$$

We have shown that:

$$\forall \boldsymbol{u} \in \mathbb{R}^{Nd}, \quad \boldsymbol{R}_T(\boldsymbol{u}) \le \frac{B_{\boldsymbol{\psi}}(\boldsymbol{A}^{1/2}\boldsymbol{u}, \boldsymbol{A}^{1/2}\boldsymbol{x}_1)}{\eta} + \eta \max_{i \le N} A_{ii}^{-1} \sum_{t=1}^{T} \frac{\|g_t\|_*^2}{2\lambda}$$

MT-OMD: Analysis

Let $A \in \mathbb{R}^{N\times N}$, $\boldsymbol{A} = A \otimes I_d \in \mathbb{R}^{Nd\times Nd}$. For a regularizer $\psi\colon \mathbb{R}^d \to \mathbb{R}$, let

$$\boldsymbol{\psi}\colon \boldsymbol{u} \in \mathbb{R}^{Nd} \mapsto \sum_{i=1}^{N} \psi(\boldsymbol{u}^{(i)}), \qquad \tilde{\boldsymbol{\psi}}\colon \boldsymbol{u} \in \mathbb{R}^{Nd} \mapsto \boldsymbol{\psi}(\boldsymbol{A}^{1/2}\boldsymbol{u})$$

We have $B_{\tilde{\boldsymbol{\psi}}}(\boldsymbol{x}, \boldsymbol{y}) = B_{\boldsymbol{\psi}}(\boldsymbol{A}^{1/2}\boldsymbol{x}, \boldsymbol{A}^{1/2}\boldsymbol{y})$, so the MT-OMD update writes:

$$\boldsymbol{x}_{t+1} = \underset{x\in\boldsymbol{V}}{\operatorname{argmin}} \ \langle \eta_t \bar{\boldsymbol{g}}_t, \boldsymbol{x}\rangle + B_{\boldsymbol{\psi}}(\boldsymbol{A}^{1/2}\boldsymbol{x}, \boldsymbol{A}^{1/2}\boldsymbol{x}_t)$$

$$= \boldsymbol{A}^{-1/2} \underset{\boldsymbol{y}\in\boldsymbol{A}^{1/2}(\boldsymbol{V})}{\operatorname{argmin}} \ \langle \eta_t \boldsymbol{A}^{-1/2}\bar{\boldsymbol{g}}_t, \boldsymbol{y}\rangle + B_{\boldsymbol{\psi}}(\boldsymbol{y}, \boldsymbol{y}_t)$$

We have shown that:

$$\forall \boldsymbol{u} \in \mathbb{R}^{Nd}, \quad \boldsymbol{R}_T(\boldsymbol{u}) \le \frac{B_{\boldsymbol{\psi}}(\boldsymbol{A}^{1/2}\boldsymbol{u}, \boldsymbol{A}^{1/2}\boldsymbol{x}_1)}{\eta} + \eta \max_{i\le N} A_{ii}^{-1} \sum_{t=1}^{T} \frac{\|g_t\|_\star^2}{2\lambda}$$

Multitask OGD (1/2)

Instantiating the previous bound for MT-OGD ($\psi = \frac{1}{2}\|\cdot\|_2^2$), we obtain:

$$\forall \boldsymbol{u} \in \mathbb{R}^{Nd}, \quad \boldsymbol{R}_T(\boldsymbol{u}) \leq \frac{(\boldsymbol{u}-\boldsymbol{x}_1)^\top \boldsymbol{A}(\boldsymbol{u}-\boldsymbol{x}_1)}{2\eta} + \eta \max_{i\leq N} A_{ii}^{-1} \sum_{t=1}^{T} \frac{\|g_t\|_2^2}{2\lambda}$$

If $A = I_N + b\left(I_N - \frac{\mathbb{1}\mathbb{1}^\top}{N}\right)$ (and $\boldsymbol{x}_1 = 0$), we obtain:

$$\begin{aligned}\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u} &= \|\boldsymbol{u}\|_2^2 + b\sum_{i=1}^{N} \|\boldsymbol{u}^{(i)} - \bar{\boldsymbol{u}}\|_2^2 \\ &= \|\boldsymbol{u}\|_2^2 + b(N-1)\,Var(\boldsymbol{u})\end{aligned}$$

and

$$\max_{i\leq N} A_{ii}^{-1} = \frac{b+N}{(1+b)N}$$

Multitask OGD (2/2)

Under which condition? Tasks have a small variance

Let

$$V = \{u \in \mathbb{R}^d \colon \|u\|_2 \le D\}$$
$$\boldsymbol{V} = \{\boldsymbol{u} \in \mathbb{R}^{Nd} \colon \|\boldsymbol{u}^{(i)}\|_2 \le D \ \ \forall i \le N\}$$
$$\boldsymbol{V}_\sigma = \{\boldsymbol{u} \in \boldsymbol{V} \colon Var(\boldsymbol{u}) \le \sigma^2 D^2\}$$

For all $\boldsymbol{u} \in \boldsymbol{V}_\sigma$ we have:

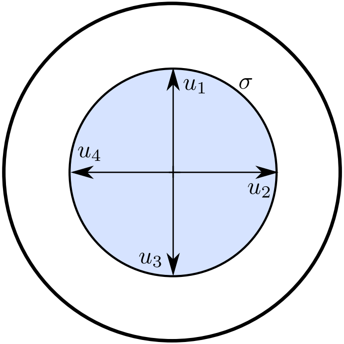
$$\forall \boldsymbol{u} \in \mathbb{R}^{Nd}, \quad \boldsymbol{R}_T(\boldsymbol{u}) \le \frac{ND^2(1 + b\frac{N-1}{N}\sigma^2)}{2\eta} + \frac{\eta(b+N)}{(1+b)N} \sum_{t=1}^{T} \frac{\|g_t\|_2^2}{2\lambda}$$
$$\le DL_g \sqrt{1 + \sigma^2(N-1)}\sqrt{2T}$$

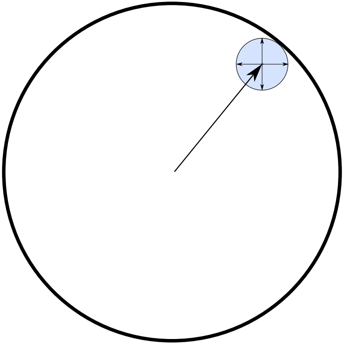
after optimizing η and b. Recall that independent OGDs give $DL_g\sqrt{NT}$.
Nicely interpolates between the extreme cases $\sigma = 0$ and $\sigma = 1$.

Matching Lower Bound / Separation Result

For any algorithm

$$\boldsymbol{R}_T \geq \frac{1}{4}\left(DL_g\sqrt{1+\sigma^2(N-1)}\sqrt{2T}\right) .$$





Extension to Any Norm

If

$$Var_{\|\cdot\|}(\boldsymbol{u}) = \frac{1}{N-1} \sum_{i=1}^{N} \left\| \boldsymbol{u}^{(i)} - \bar{\boldsymbol{u}} \right\|^2 ,$$

then

$$R_T(\boldsymbol{u}) \leq DL_g \sqrt{1 + \sigma^2(N-1)} \sqrt{8T} .$$

In particular,

$$R_T(\boldsymbol{u}) \leq L_g \sqrt{1 + \sigma^2(N-1)} \sqrt{16eT \ln d} .$$

Multitask EG (1/2)

Recall that $A = I_N + b\left(I_N - \frac{\mathbb{1}\mathbb{1}^\top}{N}\right)$.

For $\psi = \frac{1}{2}\|\cdot\|_2^2$, $\quad B_{\boldsymbol{\psi}}(\boldsymbol{A}^{1/2}\boldsymbol{u}, 0) = \sum_{i=1}^N \|\boldsymbol{u}^{(i)}\|_2^2 + b \; Var(\boldsymbol{u})$

For $\psi(x) = \sum_j x_j \ln x_j$, $\quad B_{\boldsymbol{\psi}}(\boldsymbol{A}^{1/2}\boldsymbol{u}, \frac{\mathbb{1}}{d}) \leq N \ln d$, for all $\boldsymbol{A}^{1/2}\boldsymbol{u} \in \boldsymbol{\Delta}$

Plugging and optimizing η yields for MT-EG:

$$\boldsymbol{R}_T \leq L_g \sqrt{\frac{2(b+N)}{b+1}} \sqrt{T \ln d}$$

Multitask EG (2/2)

$$\boldsymbol{R}_T \leq L_g \sqrt{\frac{2(b+N)}{b+1}} \sqrt{T \ln d}$$

But $(\boldsymbol{A}^{1/2}\boldsymbol{u})^{(i)} = \sqrt{1+b}\,\boldsymbol{u}^{(i)} + (1-\sqrt{1+b})\bar{\boldsymbol{u}}$.

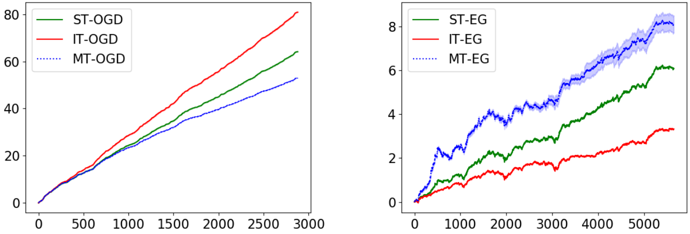
We should choose $b^* = \max\{b \geq 0 \colon \boldsymbol{A}^{1/2}\boldsymbol{u} \in \boldsymbol{\Delta}\}$.

Let $Var_\Delta(\boldsymbol{u}) = \max_{j \leq d} \left(\frac{\boldsymbol{u}_j^{max} - \boldsymbol{u}_j^{min}}{\boldsymbol{u}_j^{max}}\right)^2$. For every $\boldsymbol{u} \in \Delta$ such that $Var_\Delta(\boldsymbol{u}) \leq \sigma^2$, choosing $b = \frac{1-\sigma^2}{\sigma^2}$ yields:

$$\boldsymbol{R}_T \leq L_g \sqrt{1+\sigma^2(N-1)} \sqrt{2T \ln d}\,.$$

Experiments (1/2)

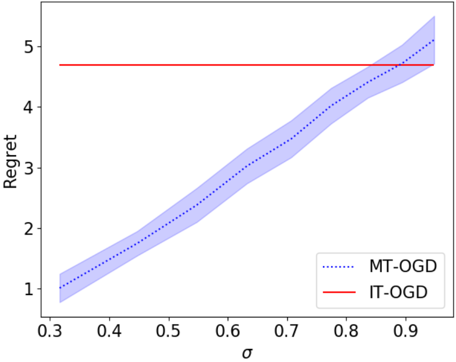
Both MT-OGD and MT-EG enjoy closed form updates. Experiments show an improvement upon both Independent Task OMD (IT-OMD, $b = 0$) and Single Task OMD (ST-OMD, $b = +\infty$).



Cumulative losses for MT-OGD on the lenk dataset (left) and cumulative wealth for MT-EG on the NYSE dataset (right).

Experiments (2/2)

Regret against task standard deviation σ (in accordance with the upper/lower bounds).



Conclusion

- MT-OMD induces the **multitask acceleration**:

$$\sqrt{1+\sigma^2(N-1)} \quad VS. \quad \sqrt{N}$$

- **How?** By sharing gradients between agents, $\tilde{\psi} = \psi(\boldsymbol{A}^{1/2}\cdot)$
- **Under which condition?** Task variance $\sigma^2 \leq 1$
- Enjoy closed form updates for MT-OGD and MT-EG
- The multitask acceleration is orthogonal to other kinds of refinements (q-norms, adaptive learning rates, smooth losses)
- Limitation: requires the knowledge of σ^2

On the choice of A

$A = (1+b)I_N - \frac{b}{N}\mathbb{1}\mathbb{1}^\top$ can actually be rewritten $A = I_N + bL^{clique}$.

If $A = I_N + bL^G$ for a generic graph G, with weight matrix W, we have:

$$\boldsymbol{u}^\top \boldsymbol{A}\boldsymbol{u} = \|\boldsymbol{u}\|_2^2 + b\sum_{i,j} W_{ij}\|\boldsymbol{u}^{(i)} - \boldsymbol{u}^{(j)}\|_2^2$$

Allows to encode more precise knowledge about the task variance.
But the computation of A_{ii}^{-1} has to be done on a case by case basis.
Works also for the variance definition on the probability simplex Δ.