Multitask Online Mirror Descent

Anonymous authors
Paper under double-blind review

Abstract

We introduce and analyze MT-OMD, a multitask generalization of Online Mirror Descent (OMD) which operates by sharing updates between tasks. We prove that the regret of MT-OMD is of order $\sqrt{1+\sigma^2(N-1)}\sqrt{T}$, where σ^2 is the task variance according to the geometry induced by the regularizer, N is the number of tasks, and T is the time horizon. Whenever tasks are similar, that is $\sigma^2 \leq 1$, our method improves upon the $\sqrt{NT}$ bound obtained by running independent OMDs on each task. We further provide a matching lower bound, and show that our multitask extensions of Online Gradient Descent and Exponentiated Gradient, two major instances of OMD, enjoy closed-form updates, making them easy to use in practice. Finally, we present experiments which support our theoretical findings.

1 Introduction

In multitask learning (Caruana, 1997), one faces a set of tasks to solve, and tries to leverage their similarities to learn faster. Task similarity is often formalized in terms of Euclidean distances among the best performing models for each task, see Evgeniou & Pontil (2004) for an example. However, in online convex optimization, and Online Mirror Descent (OMD) in particular, it is well known that using different geometries to measure distances in the model space can bring substantial advantages — see, e.g., Hazan (2016); Orabona (2019). For instance, when the model space is the probability simplex in $\mathbb{R}^d$, running OMD with the KL divergence (corresponding to an entropic regularizer) allows one to learn at a rate depending only logarithmically on d. It is thus natural to investigate to what extent measuring task similarities using geometries that are possibly non-Euclidean could improve the analysis of online multitask learning. From an application perspective, typical online multitask scenarios include federated learning applications for mobile users (e.g., personalized recommendation or health monitoring) or for smart homes (e.g., energy consumption prediction), mobile sensor networks for environmental monitoring, or even networked weather forecasting. These scenarios fit well with online learning, as new data is being generated all the time, and require different losses and decision sets, motivating the design of a general framework.

In this work, we introduce MT-OMD, a multitask generalization of OMD which applies to any strongly convex regularizer. We present a regret analysis establishing that MT-OMD outperforms OMD (run independently on each task) whenever tasks are similar according to the geometry induced by the regularizer. Our work builds on the multitask extension of the Perceptron algorithm developed in Cavallanti et al. (2010), where prior knowledge about task similarities is expressed through a symmetric positive definite interaction matrix A. Typically, $A = I + L$, where L is the Laplacian of a task relatedness graph with adjacency matrix W. The authors then show that the number of mistakes depends on $\sum_i \|u_i\|_2^2 + \sum_{i,j} W_{ij}\|u_i - u_j\|_2^2$, where each u_i denotes the best model for task i. This expression can be seen as a measure of task dispersion with respect to matrix W and norm $\|\cdot\|_2$. The Euclidean norm appears because the Perceptron is an instance of OMD for the hinge loss with the Euclidean regularizer, so that distances in the model space are measured through the corresponding Bregman divergence, which is the Euclidean squared norm.

For an arbitrary strongly convex regularizer ψ, the regret of OMD is controlled by a Bregman divergence and a term inversely proportional to the curvature of the regularizer. The key challenge we face is how to extend the OMD regularizer to the multitask setting so that the dispersion term captures task similarities. A natural strategy would be to choose a regularizer whose Bregman divergence features $\sum_{i,j} W_{ij} B_\psi(u_i, u_j)$.

Although this mimics the Euclidean dispersion term of the Perceptron, the associated regularizer has a small curvature, compromising the divergence-curvature balance which, as we said, controls the regret. Observing that the Perceptron's dispersion term can be rewritten $\|\boldsymbol{A}^{1/2}\boldsymbol{u}\|_2^2$, where $\boldsymbol{A}$ is a block version (across tasks) of A and $\boldsymbol{u}$ is the concatenation of the reference vectors u_i, our solution consists in using the regularizer $\boldsymbol{\psi}(\boldsymbol{A}^{1/2}\,\cdot\,)$, where $\boldsymbol{\psi}$ is the compound version of any base regularizer ψ defined on the model space. While exhibiting the right curvature, this regularizer has still the drawback that $\boldsymbol{A}^{1/2}\boldsymbol{u}$ might be outside the domain of $\boldsymbol{\psi}$. To get around this difficulty, we introduce a notion of variance aligned with the geometry induced by $\boldsymbol{\psi}$, such that the corresponding Bregman divergence $B_{\boldsymbol{\psi}}(\boldsymbol{A}^{1/2}\boldsymbol{u}, \boldsymbol{A}^{1/2}\boldsymbol{v})$ is always defined for sets of tasks with small variance. We then show that the Bregman divergence can be upper bounded in terms of the task variance σ^2, and by tuning appropriately the matrix A we obtain a regret bound for MT-OMD that scales as $\sqrt{1+\sigma^2(N-1)}$. In contrast, the regret of independent OMD scales as $\sqrt{N}$, highlighting the advantage brought by MT-OMD when tasks have a small variance. We stress that this improvement is independent of the chosen regularizer, thereby offering a substantial acceleration in a wide range of scenarios. To keep the exposition simple, we first work with a fixed and known σ^2. We then show an extension of MT-OMD that does not require any prior knowledge on the task similarity. The rest of the paper is organized as follows. In Section 2, we introduce the multitask online learning problem and describe MT-OMD, our multitask extension to solve it. In Section 3, we derive a regret analysis for MT-OMD, which highlights its advantage when tasks are similar. Section 4 is devoted to algorithmic implementations, and Section 5 to experiments.

Related work. Starting from the seminal work by Caruana (1997), multitask learning has been intensively studied for more than two decades, see Zhang & Yang (2021) for a recent survey. Similarly to Cavallanti et al. (2010), our work is inspired by the Laplacian multitask framework of Evgeniou et al. (2005). This framework has been extended to kernel-based learning (Sheldon, 2008), kernel-based unsupervised learning (Gu et al., 2011), contextual bandits (Cesa-Bianchi et al., 2013), spectral clustering (Yang et al., 2014), stratified model learning (Tuck et al., 2021), and, more recently, federated learning (Dinh et al., 2021). See also Herbster & Lever (2009) for different applications of Laplacians in online learning. A multitask version of OMD has been previously proposed by Kakade et al. (2012). Their approach, unlike ours, is cast in terms of matrix learning, and uses group norms and Schatten p-norm regularizers. Their bounds scale with the diameter of the model space according to these norms (as opposed to scaling with the task variance, as in our analysis). Moreover, their learning bias is limited to the choice of the matrix norm regularizer and does not explicitly include a notion of task similarity matrix. Abernethy et al. (2007); Dekel et al. (2007) investigate different multitask extensions of online learning, see also Alquier et al. (2017); Finn et al. (2019); Balcan et al. (2019); Denevi et al. (2019) for related extensions to meta-learning. Some online multitask applications are studied in Pillonetto et al. (2008); Li et al. (2014; 2019), but without providing any regret analyses. Saha et al. (2011); Zhang et al. (2018) extend the results of Cavallanti et al. (2010) to dynamically updated interaction matrices. However, no regret bounds are provided. Murugesan et al. (2016) look at distributed online classification and prove regret bounds, but they are not applicable in our asynchronous model. Other approaches for learning task similarities include Zhang & Yeung (2010); Pentina & Lampert (2017); Shui et al. (2019). We finally note the recent work by Boursier et al. (2022), which establishes multitask learning guarantees with trace norm regularization when the number of samples per task is small, and that by Laforgue et al. (2022), which learns jointly the tasks and their structure, but only with the Euclidean regularizer and under the assumption that the task activations are stochastic.

Although our asynchronous multitask setting is identical to that of Cavallanti et al. (2010), we emphasize that our work extends theirs much beyond the fact that we consider arbitrary convex losses instead of just the hinge loss. Algorithmically, MT-OMD generalizes the Multitask Perceptron in much the same way OMD generalizes the standard Perceptron. From a technical point of view, Theorem 1 in Cavallanti et al. (2010) is a direct consequence of the Kernel Perceptron Theorem, and is therefore limited to Euclidean geometries. Instead, our work provides a complete analysis of all regularizers of the form $\boldsymbol{\psi}(\boldsymbol{A}^{1/2}\cdot)$. Although Cavallanti et al. (2010) also contains a non-Euclidean p-norm extension of the Multitask Perceptron, we point out that their extension is based on a regularizer of the form $\|\boldsymbol{A}\boldsymbol{u}\|_p^2$. This is different from MT-OMD for p-norms, which instead uses $\sum_i \|(\boldsymbol{A}^{1/2}\boldsymbol{u})^{(i)}\|_p^2$. As a consequence, their bound is worse than ours (see Appendix C for technical details), does not feature any variance term, and does not specialize to the Euclidean case when $p=2$. Note that our analysis on the simplex is also completely novel as far as we know.

2 Multitask Online Learning

We now describe the multitask online learning problem, and introduce our approach to solve it. We use a cooperative and asynchronous multiagent formalism: the online algorithm is run in a distributed fashion by communicating agents, that however make predictions at different time steps.

Problem formulation and reminders on OMD. We consider an online convex optimization setting with a set of $N \in \mathbb{N}$ agents, each learning a possibly different task on a common convex decision set $V \subset \mathbb{R}^d$. At each time step $t = 1, 2, \dots$ some agent $i_t \leq N$ makes a prediction $x_t \in V$ for its task, incurs loss $\ell_t(x_t)$, and observes a subgradient of ℓ_t at x_t, where ℓ_t is a convex loss function. We say that i_t is the active agent at time t. Both the sequence $i_1, i_2, \dots$ of active agents and the sequence $\ell_1, \ell_2, \dots$ of convex losses are chosen adversarially and hidden from the agents. Our goal is to minimize the *multitask regret*, which is defined as the sum of the individual regrets

$$R_T = \sum_{i=1}^{N} \left(\sum_{t\colon i_t = i} \ell_t(x_t) - \inf_{u \in V} \sum_{t\colon i_t = i} \ell_t(u) \right) = \sum_{t=1}^{T} \ell_t(x_t) - \sum_{i=1}^{N} \inf_{u \in V} \sum_{t\colon i_t = i} \ell_t(u)\,. \tag{1}$$

A natural idea to minimize Equation (1) is to run N independent OMDs, one for each agent. Recall that OMD refers to a family of algorithms, typically used to minimize a regret of the form $\sum_t \ell_t(x_t) - \inf_{u \in V} \sum_t \ell_t(u)$, for any sequence of proper convex loss functions ℓ_t. An instance of OMD is parameterized by a λ-strongly convex regularizer $\psi\colon \mathbb{R}^d \to \mathbb{R}$, and has the update rule

$$x_{t+1} = \operatorname*{arg\,min}_{x \in V} \quad \langle \eta_t g_t, x \rangle + B_\psi(x, x_t)\,, \tag{2}$$

where $g_t \in \mathbb{R}^d$ is a subgradient of ℓ_t at point x_t, $B_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y \rangle$ denotes the Bregman divergence associated to ψ, and $\eta_t > 0$ is a tunable learning rate. Standard results allow to bound the regret achieved by the sequence of iterates produced by OMD. For a fixed η and any initial point $x_1 \in V$, we have (Orabona, 2019, Theorem 6.8) that for all $u \in V$

$$\sum_{t=1}^{T} \ell_t(x_t) - \ell_t(u) \leq \frac{B_\psi(u, x_1)}{\eta} + \frac{\eta}{2\lambda} \sum_{t=1}^{T} \|g_t\|_\star^2\,, \tag{3}$$

with $\|\cdot\|_\star$ the dual norm of the norm with respect to which ψ is strongly convex (see Definition 4 in the Appendix). The choice of the regularizer ψ shapes the above bound through the quantities $B_\psi(u, x_1)$ and $\|g_t\|_\star$. When $\psi = \frac{1}{2}\|\cdot\|_2^2$, we have $B_\psi(x, y) = \frac{1}{2}\|x - y\|_2^2$, $\|\cdot\|_\star = \|\cdot\|_2$, $\lambda = 1$, and the algorithm is called Online Gradient Descent (OGD). However, depending on the problem, a different choice of the regularizer might better captures the underlying geometry. A well-known example is Exponentiated Gradient (EG), an instance of OMD in which V is the probability simplex in $\mathbb{R}^d$, such that $V = \Delta \coloneqq \{x \in \mathbb{R}_+^d\colon \sum_j x_j = 1\}$. EG uses the negative entropy regularizer $x \mapsto \sum_j x_j \ln(x_j)$, and assuming that $\|g_t\|_\infty \leq L_g$, one achieves bounds of order $\mathcal{O}(L_g\sqrt{T \ln d})$, while OGD yields bounds of order $\mathcal{O}(L_g\sqrt{dT})$. We emphasize that our cooperative extension adapts to several types of regularizers, and can therefore exploit these improvements with respect to the dependence on d, see Proposition 8. Let C be a generic constant such that $C\sqrt{T}$ bounds the regret incurred by the chosen OMD (e.g., $C = L_g\sqrt{\ln d}$, or $C = L_g\sqrt{d}$ above). Then, by Jensen's inequality the multitask regret of N independent OMDs satisfies

$$R_T \leq \sum_{i=1}^{N} C\sqrt{T_i} \leq C\sqrt{NT}\,, \tag{4}$$

where $T_i = \sum_{t=1}^{T} \mathbb{I}\{i_t = i\}$ denotes the number of times agent i was active. Our goal is to show that introducing communication between the agents may significantly improve on Equation (4) with respect to the dependence on N.

A multitask extension. We now describe our multitask OMD approach. To gain some insights on it, we first focus on OGD. For $i \leq N$ and $t \leq T$, let $x_{i,t} \in \mathbb{R}^d$ denote the prediction maintained by agent i at time step t. By completing the square in Equation (2) for $\psi = \psi_{\text{Euc}} \coloneqq \frac{1}{2}\|\cdot\|_2^2$, the independent OGDs updates can be rewritten for all $i \leq N$ and t such that $i_t = i$:

$$x_{i,t+1} = \Pi_{V,\|\cdot\|_2}\big(x_{i,t} - \eta_t g_t\big)\,, \tag{5}$$

where $\Pi_{V,\|\cdot\|}$ denotes the projection operator onto the convex set V according to the norm $\|\cdot\|$, that is $\Pi_{V,\|\cdot\|}(x) = \arg\min_{y\in V}\|x-y\|$. Our analysis relies on *compound representations*, that we explain next. We use bold notation to refer to compound vectors, such that for $u_1,\dots,u_N \in \mathbb{R}^d$, the compound vector is $\boldsymbol{u} = [u_1,\dots,u_N] \in \mathbb{R}^{Nd}$. For $i \leq N$, we use $\boldsymbol{u}^{(i)}$ to refer to the i^{th} block of $\boldsymbol{u}$, such that $\boldsymbol{u}^{(i)} = u_i$ in the above example. So $\boldsymbol{x}_t$ is the compound vector of the $(x_{i,t})_{i=1}^N$, such that $x_t = \boldsymbol{x}_t^{(i_t)}$, and the multitask regret rewrites as $R_T(\boldsymbol{u}) = \sum_{t=1}^T \ell_t\big(\boldsymbol{x}_t^{(i_t)}\big) - \ell_t\big(\boldsymbol{u}^{(i_t)}\big)$. For any set $V \subset \mathbb{R}^d$, let $\boldsymbol{V} = V^{\otimes N} \subset \mathbb{R}^{Nd}$ denote the compound set such that $\boldsymbol{u} \in \boldsymbol{V}$ is equivalent to $\boldsymbol{u}^{(i)} \in V$ for all $i \leq N$. Equipped with this notation, the independent OGD updates Equation (5) rewrite as

$$\boldsymbol{x}_{t+1} = \Pi_{\boldsymbol{V},\|\cdot\|_2}\big(\boldsymbol{x}_t - \eta_t \boldsymbol{g}_t\big)\,, \tag{6}$$

with $\boldsymbol{g}_t \in \mathbb{R}^{Nd}$ such that $\boldsymbol{g}_t^{(i)} = g_t$ for $i = i_t$, and $0_{\mathbb{R}^d}$ otherwise. In other words, only the active agent has a non-zero gradient and therefore makes an update. Our goal is to incorporate communication into this independent update. To that end, we consider the general idea of *sharing updates* by considering (sub) gradients of the form $\boldsymbol{A}^{-1}\boldsymbol{g}_t$, where $\boldsymbol{A}^{-1} \in \mathbb{R}^{Nd\times Nd}$ is a shortcut notation for $A^{-1}\otimes I_d$ and $A \in \mathbb{R}^{N\times N}$ is any symmetric positive definite interaction matrix. Note that A is a parameter of the algorithm playing the role of a learning bias. While our central result (Theorem 1) holds for any choice of A, our more specialized bounds (see Propositions 5 to 8) apply to a parameterized family of matrices A. A simple computation shows that $(\boldsymbol{A}^{-1}\boldsymbol{g}_t)^{(i)} = A^{-1}_{ii_t} g_t$. Thus, every agent i makes an update proportional to $A^{-1}_{ii_t}$ at each time step t. In other words, the active agent (the only one to suffer a loss) shares its update with the other agents. Results in Section 3 are proved by designing a matrix A^{-1} (or equivalently A) such that $A^{-1}_{ii_t}$ captures the similarity between tasks i and i_t. Intuitively, the active agent i_t should share its update (gradient) with another agent i to the extent their respective tasks are similar. Overall, denoting by $\|u\|_M = \sqrt{u^\top M u}$ the Mahalanobis norm of u, the MT-OGD update writes

$$\boldsymbol{x}_{t+1} = \Pi_{\boldsymbol{V},\|\cdot\|_{\boldsymbol{A}}}\big(\boldsymbol{x}_t - \eta_t \boldsymbol{A}^{-1}\boldsymbol{g}_t\big)\,. \tag{7}$$

In comparison to Equation (6), the need for changing the norm in the projection, although unclear at first sight, can be explained in multiple ways. First, it is key to the analysis, as we see in the proof of Theorem 1. Second, it can be interpreted as another way of exchanging information between agents, see Remark 1. Finally, note that update Equation (7) can be decomposed as

$$\begin{aligned}
\widetilde{\boldsymbol{x}}_{t+1} &= \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^{Nd}} \ \langle \eta_t \boldsymbol{g}_t, \boldsymbol{x}\rangle + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}_t\|_{\boldsymbol{A}}^2\,,\\
\boldsymbol{x}_{t+1} &= \operatorname*{arg\,min}_{\boldsymbol{x}\in\boldsymbol{V}} \ \frac{1}{2}\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}_{t+1}\|_{\boldsymbol{A}}^2\,,
\end{aligned} \tag{8}$$

showing that it is natural to keep the same norm in both updates. Most importantly, what Equation (8) tells us, is that the MT-OGD update rule is actually an OMD update—see e.g., (Orabona, 2019, Section 6.4)—with regularizer $\boldsymbol{x} \mapsto \frac{1}{2}\|\boldsymbol{x}\|_{\boldsymbol{A}}^2 = \boldsymbol{\psi}_{\text{Euc}}\big(\boldsymbol{A}^{1/2}\boldsymbol{x}\big)$. This provides a natural path for extending our multitask approach to any regularizer. Given a base regularizer $\psi\colon \mathbb{R}^d \to \mathbb{R}$, the *compound regularizer* $\boldsymbol{\psi}$ is given by $\boldsymbol{\psi}\colon \boldsymbol{x} \in \mathbb{R}^{Nd} \mapsto \sum_{i=1}^N \psi\big(\boldsymbol{x}^{(i)}\big)$. When there exists a function $\phi\colon \mathbb{R}\to\mathbb{R}$ such that $\psi(x) = \sum_j \phi(x_j)$, the compound regularizer is the natural extension of ψ to $\mathbb{R}^{Nd}$. Note, however, that the relationship can be more complex, e.g., when $\psi(x) = \frac{1}{2}\|x\|_p^2$. Using regularizer $\boldsymbol{\psi}\big(\boldsymbol{A}^{1/2}\cdot\big)$, whose associate divergence is $B_{\boldsymbol{\psi}}\big(\boldsymbol{A}^{1/2}\boldsymbol{x}, \boldsymbol{A}^{1/2}\boldsymbol{x}'\big)$, the MT-OMD update thus reads

$$\boldsymbol{x}_{t+1} = \operatorname*{arg\,min}_{\boldsymbol{x}\in\boldsymbol{V}} \ \langle \eta_t \boldsymbol{g}_t, \boldsymbol{x}\rangle + B_{\boldsymbol{\psi}}\big(\boldsymbol{A}^{1/2}\boldsymbol{x}, \boldsymbol{A}^{1/2}\boldsymbol{x}_t\big)\,. \tag{9}$$

Clearly, if $\psi = \psi_{\text{Euc}}$, we recover the MT-OGD update. Observe also that whenever $A = I_N$, MT-OMD is equivalent to N independent OMDs. We conclude this exposition with a remark shedding light on the way MT-OMD introduces communication between agents.

Remark 1. *Denoting $\boldsymbol{y}_t = \boldsymbol{A}^{1/2}\boldsymbol{x}_t$, Equation (9) rewrites*

$$\boldsymbol{x}_{t+1} = \boldsymbol{A}^{-1/2} \operatorname*{arg\,min}_{\boldsymbol{y} \in \boldsymbol{A}^{1/2}(\boldsymbol{V})} \langle \eta_t \boldsymbol{A}^{-1/2}\boldsymbol{g}_t, \boldsymbol{y}\rangle + B_{\boldsymbol{\psi}}(\boldsymbol{y}, \boldsymbol{y}_t)\,. \tag{10}$$

The two occurrences of $\boldsymbol{A}^{-1/2}$ reveal that agents communicate in two distinct ways: one through the shared update (the innermost occurrence of $\boldsymbol{A}^{-1/2}$), and one through computing the final prediction $\boldsymbol{x}_{t+1}$ as a linear combination of the solution to the optimization problem. Multiplying Equation (10) by $\boldsymbol{A}^{1/2}$, MT-OMD can also be seen as standard OMD on the transformed iterate $\boldsymbol{y}_t$.

3 Regret Analysis

We now provide a regret analysis for MT-OMD. We start with a general theorem presenting two bounds, for constant and time-varying learning rates. These results are then instantiated to different types of regularizer and variance in Propositions 2 to 8. The main difficulty is to characterize the strong convexity of $\boldsymbol{\psi}\big(\boldsymbol{A}^{1/2}\,\cdot\,\big)$, see Lemmas 10 and 11 in the Appendix. Throughout the section, $V \subset \mathbb{R}^d$ is a convex set of comparators, and $(\ell_t)_{t=1}^T$ is a sequence of proper convex loss functions chosen by the adversary. Note that all technical proofs can be found in Appendix A.

Theorem 1. *Let $\psi\colon \mathbb{R}^d \to \overline{\mathbb{R}}$ be λ-strongly convex with respect to norm $\|\cdot\|$ on V, let $A \in \mathbb{R}^{N\times N}$ be symmetric positive definite, and set $\boldsymbol{x}_1 \in \boldsymbol{V}$. Then, MT-OMD with $\eta_t \coloneqq \eta$ produces a sequence of iterates $(\boldsymbol{x}_t)_{t=1}^T$ such that for all $\boldsymbol{u} \in \boldsymbol{V}$, $R_T(\boldsymbol{u})$ is bounded by*

$$\frac{B_{\boldsymbol{\psi}}\big(\boldsymbol{A}^{1/2}\boldsymbol{u}, \boldsymbol{A}^{1/2}\boldsymbol{x}_1\big)}{\eta} \;+\; \max_{i\le N} A_{ii}^{-1}\; \frac{\eta}{2\lambda}\sum_{t=1}^T \|g_t\|_\star^2\,. \tag{11}$$

Moreover, for any sequence of nonincreasing learning rates $(\eta_t)_{t=1}^T$, MT-OMD produces a sequence of iterates $(\boldsymbol{x}_t)_{t=1}^T$ such that for all $\boldsymbol{u} \in \boldsymbol{V}$, $R_T(\boldsymbol{u})$ is bounded by

$$\max_{t\le T} \frac{B_{\boldsymbol{\psi}}\big(\boldsymbol{A}^{1/2}\boldsymbol{u}, \boldsymbol{A}^{1/2}\boldsymbol{x}_t\big)}{\eta_T} \;+\; \max_{i\le N} A_{ii}^{-1}\; \frac{1}{2\lambda}\sum_{t=1}^T \eta_t\|g_t\|_\star^2\,. \tag{12}$$

3.1 Multitask Online Gradient Descent

For $\psi = \frac{1}{2}\|\cdot\|_2^2$, $A = I_N$ (independent updates), unit-norm reference vectors $(\boldsymbol{u}^{(i)})_{i=1}^N$, L_g-Lipschitz losses, and $\boldsymbol{x}_1 = \boldsymbol{0}$, bound Equation (11) becomes: $ND^2/2\eta + \eta T L_g^2/2$. Choosing $\eta = D\sqrt{N}/L_g\sqrt{T}$, we recover the $DL_g\sqrt{NT}$ bound of Equation (4). Our goal is to design interaction matrices A that make Equation (11) smaller. In the absence of additional assumptions on the set of comparators, it is however impossible to get a systematic improvement: the bound is a sum of two terms, and introducing interactions typically reduces one term but increases the other. To get around this difficulty, we introduce a simple condition on the task similarity, that allows us to control the increase of $B_{\boldsymbol{\psi}}\big(\boldsymbol{A}^{1/2}\boldsymbol{u}, \boldsymbol{A}^{1/2}\boldsymbol{x}_1\big)$ for a carefully designed class of interaction matrices.

Definition 1. *Let $\|\cdot\|\colon \mathbb{R}^d \to \mathbb{R}$ be any norm, and $\bar{\boldsymbol{u}} = (1/N)\sum_{i=1}^N \boldsymbol{u}^{(i)}$, for any $\boldsymbol{u} \in \mathbb{R}^{Nd}$. We define the variance of $\boldsymbol{u}$ w.r.t. $\|\cdot\|$ as*

$$\operatorname{Var}_{\|\cdot\|}(\boldsymbol{u}) = \frac{1}{N-1}\sum_{i=1}^N \big\|\boldsymbol{u}^{(i)} - \bar{\boldsymbol{u}}\big\|^2\,.$$

Let $D = \sup_{u\in V}\|u\|$, and $\sigma > 0$. The comparators with variance smaller than $\sigma^2 D^2$ are denoted by

$$\boldsymbol{V}_{\|\cdot\|,\sigma} = \big\{\boldsymbol{u} \in \boldsymbol{V}\colon \operatorname{Var}_{\|\cdot\|}(\boldsymbol{u}) \le \sigma^2 D^2\big\}\,. \tag{13}$$

For sets of comparators of the form Equation (13), we show that MT-OGD achieves significant improvements over its independent counterpart. The rationale behind this gain is fairly natural: the tasks associated with comparators in Equation (13) are similar due to the variance constraint, so that communication indeed helps. Note that condition Equation (13) does not enforce any restriction on the norms of the individual $\boldsymbol{u}^{(i)}$, and is much more complex than a simple rescaling of the feasible set by σ^2. For instance, one could imagine task vectors highly concentrated around some vector u_0, whose norm is D: the individual norms are close to D, but the task variance is small. This is precisely the construction used in the separation result (Proposition 4). As MT-OMD leverages the additional information of the task variance (unavailable in the independent case), it is expected that an improvement *should be* possible. The problems of how to use this extra information and what improvement can be achieved through it are addressed in the rest of this section. To that end, we first assume σ^2 to be known. This assumption can be seen as a learning bias, analog to the knowledge of the diameter D in standard OGD bounds. In Section 3.4, we then detail a Hedge-based extension of MT-OGD that does not require the knowledge of σ^2 and only suffers an additional regret of order $\sqrt{T \log N}$.

The class of interaction matrices we consider is defined as follows. Let $L = I_N - \mathbb{1}\mathbb{1}^\top/N$. We consider matrices of the form $A(b) = I_N + b\,L$, where $b \geq 0$ quantifies the magnitude of the communication. For more intuition about this choice, see Section 3.4. We can now state a first result highlighting the advantage brought by MT-OGD.

Proposition 2. *Let $\psi = \frac{1}{2}\|\cdot\|_2^2$, $D = \sup_{x\in V}\|x\|_2$, and $\sigma \leq 1$. Assume that $\|\partial \ell_t(x)\|_2 \leq L_g$ for all $t \leq T$ and any $x \in V$. Set $b = N$, $\boldsymbol{x}_1 = \mathbf{0}$, and $\eta = D\sqrt{N(N+1)(1+(N-1)\sigma^2)}/L_g\sqrt{2T}$. Then, MT-OGD produces a sequence of iterates $(\boldsymbol{x}_t)_{t=1}^T$ such that for all $\boldsymbol{u} \in \boldsymbol{V}_{\|\cdot\|_2,\sigma}$*

$$R_T(\boldsymbol{u}) \leq DL_g\sqrt{1+\sigma^2(N-1)}\sqrt{2T}\,. \tag{14}$$

Proof sketch. With $\psi = \frac{1}{2}\|\cdot\|_2^2$, and $\boldsymbol{x}_1 = \mathbf{0}$, we have $2B_{\boldsymbol{\psi}}\big(\boldsymbol{A}(b)^{1/2}\boldsymbol{u}, \boldsymbol{A}(b)^{1/2}\mathbf{0}\big) = \|\boldsymbol{u}\|_2^2 + b(N-1)\mathrm{Var}_{\|\cdot\|_2}(\boldsymbol{u})$, which is smaller than $ND^2(1+b\frac{N-1}{N}\sigma^2)$. Then, it is easy to check that $\big[A(b)^{-1}\big]_{ii} = \frac{b+N}{(1+b)N}$ for all $i \leq N$. Substituting these values into Equation (11), we obtain

$$R_T(\boldsymbol{u}) \leq \frac{ND^2(1+b\frac{N-1}{N}\sigma^2)}{2\eta} + \frac{\eta T L_g^2}{2}\frac{b+N}{(1+b)N}\,.$$

Finally, set $\eta = \frac{ND}{L_g}\sqrt{\frac{\left(1+b\frac{N-1}{N}\sigma^2\right)(1+b)}{(b+N)T}}$ and $b = N$. □

Thus, MT-OGD enjoys a $\sqrt{1+\sigma^2(N-1)}$ dependence, which is smaller than $\sqrt{N}$ when tasks have a variance smaller than 1. When $\sigma = 0$ (all tasks are equal), MT-OGD scales as if there were only one task. When $\sigma \geq 1$, the analysis suggests to choose $b = 0$, i.e., $A = I_N$, and one recovers the performance of independent OGDs. Note that the additional $\sqrt{2}$ factor in Equation (14) can be removed for limit cases through a better optimization in b: the bound obtained in the proof actually reads $DL\sqrt{F(\sigma)\,T}$, with $F(0) = 1$ and $F(1) = N$. However, the function F lacks of interpretability outside of the limit cases (for details see Appendix A.2) motivating our choice to present the looser but more interpretable bound Equation (14). For a large N, we have $\sqrt{1+\sigma^2(N-1)} \approx \sigma\sqrt{N}$. The improvement brought by MT-OGD is thus roughly proportional to the square root of the task variance. From now on, we refer to this gain as the *multitask acceleration*. This improvement achieved by MT-OGD is actually optimal up to constants, as revealed by the following lower bound, which is only 1/4 of Equation (14).

Proposition 3. *Under the conditions of Proposition 2, the regret of any algorithm satisfies*

$$\sup_{\boldsymbol{u}\in\boldsymbol{V}_{\|\cdot\|_2,\sigma}} R_T(\boldsymbol{u}) \geq \frac{1}{4}\left(DL_g\sqrt{1+\sigma^2(N-1)}\sqrt{2T}\right)\,.$$

Another way to gain intuition about Equation (14) is to compare it to the lower bound for OGD considering independent tasks (IT-OGD). The following separation result shows that MT-OGD may strictly improve over IT-OGD.

Proposition 4. *Let $d \geq 9$, $N = 2d$, and $\sigma \leq 1$ to be tuned later. Then, there exists $\boldsymbol{u} \in \boldsymbol{V}_{\|\cdot\|_2,\sigma}$ such that*

$$R_T^{\mathrm{IT-OGD}}(\boldsymbol{u}) \geq \frac{\sqrt{(1-2\sigma^2)N}}{4}\sqrt{2T}\,.$$

Proposition 2 then yields that for any $\sigma^2 < \frac{N-16}{18N-16}$

$$R_T^{\mathrm{IT-OGD}}(\boldsymbol{u}) > R_T^{\mathrm{MT-OGD}}(\boldsymbol{u})\,.$$

3.2 Extension to any Norm Regularizers

A natural question is: *can the multitask acceleration be achieved with other regularizers?* Indeed, the proof of Proposition 2 crucially relies on the fact that the Bregman divergence can be exactly expressed in terms of $\|\boldsymbol{u}\|_2^2$ and $\mathrm{Var}_{\|\cdot\|_2}(\boldsymbol{u})$. In the following proposition, we show that such an improvement is also possible for all regularizers of the form $\frac{1}{2}\|\cdot\|^2$, for arbitrary norms $\|\cdot\|$, up to an additional multiplicative constant. A crucial application is the use of the p-norm on the probability simplex, which is known to exhibit a logarithmic dependence in d for a well-chosen p.

Proposition 5. *Let $\|\cdot\|\colon \mathbb{R}^d \to \mathbb{R}$ be any norm, $\psi = \frac{1}{2}\|\cdot\|^2$, $D = \sup_{x\in V}\|x\|$, and $\sigma \leq 1$. Assume that $\|\partial \ell_t(x)\|_\star \leq L_g$ for all $t \leq T$, $x \in V$. Set $b = N$, $\boldsymbol{x}_1 = \boldsymbol{0}$ and $\eta = D\sqrt{N(N+1)(1+(N-1)\sigma^2)}/L_g\sqrt{2T}$. Then, MT-OMD produces a sequence of iterates $(\boldsymbol{x}_t)_{t=1}^T$ such that for all $\boldsymbol{u} \in \boldsymbol{V}_{\|\cdot\|,\sigma}$*

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

In particular, for $d \geq 3$ and $V = \Delta$, choosing $\|\cdot\| = \|\cdot\|_p$, for $p = 2\ln d/(2\ln d - 1)$, and assuming that $\|\partial \ell_t(x)\|_\infty \leq L_g$, it holds for all $\boldsymbol{u} \in \boldsymbol{\Delta}_{\|\cdot\|_p,\sigma}$

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

In comparison, under the same assumptions, bound Equation (14) would write as: $L_g\sqrt{1+\sigma^2(N-1)}\sqrt{2Td}$.

Projecting onto $\boldsymbol{V}_\sigma$. Propositions 2 and 5 reveal that whenever tasks are similar (i.e., whenever $\boldsymbol{u} \in \boldsymbol{V}_\sigma$), then using the regularizer $\psi(\boldsymbol{A}^{1/2}\cdot)$ with $A \neq I_N$ accelerates the convergence. However, this is not the only way to leverage the small variance condition. For instance, one may also use this information to directly project onto $\boldsymbol{V}_\sigma \subset \boldsymbol{V}$, by considering the update

$$\boldsymbol{x}_{t+1} = \operatorname*{arg\,min}_{\boldsymbol{x}\in\boldsymbol{V}_\sigma} \ \langle \eta_t \boldsymbol{g}_t, \boldsymbol{x}\rangle + B_\psi(\boldsymbol{A}^{1/2}\boldsymbol{x}, \boldsymbol{A}^{1/2}\boldsymbol{x}_t)\,. \tag{15}$$

Although not necessary in general (Propositions 2 and 5 show that communicating the gradients is sufficient to get an improvement), this refinement presents several advantages. First, it might be simpler to compute in practice, see Section 4. Second, it allows for adaptive learning rates, that preserve the guarantees while being independent from the horizon T (Proposition 6). Finally, it allows to derive $L^\star$ bounds with the multitask acceleration for smooth loss functions (Proposition 7). Results are stated for arbitrary norms, but bounds sharper by a factor 2 can be obtained for $\|\cdot\|_2$.

Proposition 6. *Let $\|\cdot\|\colon \mathbb{R}^d \to \mathbb{R}$ be any norm, $\psi = \frac{1}{2}\|\cdot\|^2$, $D = \sup_{x\in V}\|x\|$, and $\sigma \leq 1$. Set $b = N$, and $\eta_t = D\sqrt{N(N+1)(1+(N-1)\sigma^2)}(\sum_{i=1}^t \|g_i\|_\star^2)^{-1/2}$. Then, Equation (15) produces a sequence of iterates $(\boldsymbol{x}_t)_{t=1}^T$ such that for all $\boldsymbol{u} \in \boldsymbol{V}_{\|\cdot\|,\sigma}$*

$$R_T(\boldsymbol{u}) \leq 8D\sqrt{1+\sigma^2(N-1)}\left(\sum_{t=1}^T \|g_t\|_\star^2\right)^{1/2}.$$

Proposition 7. *Let $\|\cdot\|\colon \mathbb{R}^d \to \mathbb{R}$ be any norm, $\psi = \frac{1}{2}\|\cdot\|^2$, $D = \sup_{x\in V}\|x\|$, and $\sigma \leq 1$. Assume that the ℓ_t are M-smooth, i.e., $\|\nabla\ell_t(x) - \nabla\ell_t(y)\|_\star \leq M\|x-y\|$ for all $t \leq T$, and any $x, y \in V$. Set b and η_t as in Proposition 6. Then, update Equation (15) produces a sequence of iterates $(\boldsymbol{x}_t)_{t=1}^T$ such that for all $\boldsymbol{u} \in \boldsymbol{V}_{\|\cdot\|,\sigma}$*

$$R_T(\boldsymbol{u}) \leq 16D\sqrt{1+\sigma^2(N-1)}\left(2MD\sqrt{1+\sigma^2(N-1)} + \sqrt{M\sum_{t=1}^T \ell_t(\boldsymbol{u}^{(i_t)})}\right).$$

3.3 Regularizers on the Simplex

As seen in Propositions 2 to 7, MT-OMD induces a multitask acceleration in a wide range of settings, involving different regularizers (Euclidean norm, p-norms) and various kind of loss functions (Lipschitz continuous, smooth continuous gradients). This systematic gain suggests that multitask acceleration essentially derives from our approach, and is completely orthogonal to the improvements achievable by choosing the regularizer appropriately. Bounds combining both benefits are actually derived in the second claim of Proposition 5. However, all regularizers studied so far share a crucial feature: they are defined on the entire space $\mathbb{R}^d$. As a consequence, the divergence $B_{\boldsymbol{\psi}}(\boldsymbol{A}^{1/2}\boldsymbol{u}, \boldsymbol{A}^{1/2}\boldsymbol{x})$ is always well defined, which might not be true in general, for instance when the comparator set studied is the probability simplex Δ. A workaround consists in assigning the value $+\infty$ to the Bregman divergence whenever either of the arguments is outside of the compound simplex $\boldsymbol{\Delta} = \Delta^{\otimes N}$. The choice of the interaction matrix A then becomes critical to prevent the bound from exploding, and calls for a new definition of the variance. Indeed, note that for $i \leq N$ we have $(\boldsymbol{A}(b)^{1/2}\boldsymbol{u})^{(i)} = \sqrt{1+b}\ \boldsymbol{u}^{(i)} + (1-\sqrt{1+b})\bar{\boldsymbol{u}}$. If all $\boldsymbol{u}^{(i)}$ are equal (say to $u_0 \in \Delta$), then all $(\boldsymbol{A}(b)^{1/2}\boldsymbol{u})^{(i)}$ are also equal to u_0 and $\boldsymbol{A}(b)^{1/2}\boldsymbol{u} \in \boldsymbol{\Delta}$. However, if they are different, by definition of $\bar{\boldsymbol{u}}$, for all $j \leq d$, there exists $i \leq N$ such that $\boldsymbol{u}_j^{(i)} \leq \bar{\boldsymbol{u}}_j$. Then, for b large enough, $\sqrt{1+b}\ \boldsymbol{u}_j^{(i)} + (1-\sqrt{1+b})\bar{\boldsymbol{u}}_j$ becomes negative, and $(\boldsymbol{A}(b)^{1/2}\boldsymbol{u})^{(i)}$ is out of the simplex. Luckily, the maximum acceptable value for b can be easily deduced from the following variance definition.

Definition 2. *Let $\boldsymbol{u} \in \mathbb{R}^d$. For all $j \leq d$, let*

$$\boldsymbol{u}_j^{\max} = \max_{i \leq N} \boldsymbol{u}_j^{(i)}, \quad \textit{and} \quad \boldsymbol{u}_j^{\min} = \min_{i \leq N} \boldsymbol{u}_j^{(i)}\,.$$

Then, with the convention $0/0 = 0$ we define

$$\mathrm{Var}_{\Delta}(\boldsymbol{u}) = \max_{j \leq d} \left(\frac{\boldsymbol{u}_j^{\max} - \boldsymbol{u}_j^{\min}}{\boldsymbol{u}_j^{\max}} \right)^2,$$

and for any $\sigma \leq 1$

$$\boldsymbol{\Delta}_\sigma = \left\{ \boldsymbol{u} \in \boldsymbol{\Delta} \colon \mathrm{Var}_{\Delta}(\boldsymbol{u}) \leq \sigma^2 \right\}.$$

Equipped with this new variance definition, we can now analyze regularizers defined on the simplex.

Proposition 8. *Let $\psi : \Delta \to \mathbb{R}$ be λ-strongly convex w.r.t. norm $\|\cdot\|$, and such that there exist $x^* \in \Delta$ and $C < +\infty$ such that for all $x \in \Delta, B_\psi(x, x^*) \leq C$. Let $\sigma \leq 1$, and assume that $\|\partial \ell_t(x)\|_\star \leq L_g$ for all $t \leq T$ and $x \in \Delta$. Set $b = (1-\sigma^2)/\sigma^2$, $\boldsymbol{x}_1 = [x^*, \dots, x^*]$, and $\eta = N\sqrt{2\lambda(1+b)C}/L_g\sqrt{(b+N)T}$. Then, MT-OMD produces a sequence of iterates $(\boldsymbol{x}_t)_{t=1}^T$ such that for all $\boldsymbol{u} \in \boldsymbol{\Delta}_\sigma$*

$$R_T(\boldsymbol{u}) \leq L_g\sqrt{1+\sigma^2(N-1)}\sqrt{2CT/\lambda}\,.$$

For the negative entropy we have $x^* = \mathbb{1}/d$ and $C = \ln d$. With subgradients satisfying $\|\partial \ell_t(x)\|_\infty \leq L_g$ we obtain

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

Proposition 8 shows that the multitask acceleration is not an artifact of the Euclidean geometry, but rather a general feature of MT-OMD, as long as the variance definition is aligned with the geometry of the problem.

3.4 Adaptivity to the Task Variance

Most of the results we presented so far require the knowledge of the task variance σ^2. We now present an Hedge-based extension of MT-OMD, denoted Hedge-MT-OMD, that does not require any prior information on σ^2. First, note that for $\sigma^2 \geq 1$, MT-OMD becomes equivalent to independent OMDs. A simple approach consists then in using Hedge—see, e.g., (Orabona, 2019, Section 6.8)—over a set of experts, each running an instance of MT-OMD with a different value of σ^2 chosen on a uniform grid of the interval $[0, 1]$.[1] We can show that Hedge-MT-OGD only suffers an additional regret of order $\sqrt{T \log N}$ against MT-OGD run with the exact knowledge of $\mathrm{Var}_{\|\cdot\|_2}(\boldsymbol{u})$.

[1] Note that Hedge-MT-OGD computes the loss subgradient at arbitrary points (corresponding to the expert's predictions).

Theorem 9. *Let $D = \sup_{x \in V} \|x\|_2$, and assume that $\|\partial \ell_t(x)\|_2 \leq L_g$ for all $t \leq T$, $x \in V$. Then, for all $\boldsymbol{u} \in \boldsymbol{V}$ the regret of Hedge-MT-OGD is bounded by*

$$DL_g\left(2 + \sqrt{\log N} + \sqrt{\min\left\{\mathrm{Var}_{\|\cdot\|_2}(\boldsymbol{u}), 1\right\} \cdot N}\right)\sqrt{2T}\,.$$

Variance definition and choice of A. Note that we have $(N-1)\mathrm{Var}_{\|\cdot\|_2}(\boldsymbol{u}) = \frac{1}{N}\sum_{i,j} \|\boldsymbol{u}^{(i)} - \boldsymbol{u}^{(j)}\|_2^2 = \boldsymbol{u}^\top \boldsymbol{L}\boldsymbol{u}$, where $L = I_N - \mathbb{1}\mathbb{1}^\top/N$ is the Laplacian of the weighted clique graph over $\{1, \dots, N\}$, with edges of $1/N$. A natural extension then consists in considering variances of the form

$$\mathrm{Var}^W_{\|\cdot\|_2}(\boldsymbol{u}) = \sum_{i,j=1}^N W_{ij}\|\boldsymbol{u}^{(i)} - \boldsymbol{u}^{(j)}\|_2^2 = \boldsymbol{u}^\top \boldsymbol{L}^W \boldsymbol{u}$$

for any adjacency matrix W and its Laplacian L^W. For instance, if we expect tasks to be concentrated in clusters, it is natural to consider $W_{ij} = 1$ if $\boldsymbol{u}^{(i)}$ and $\boldsymbol{u}^{(j)}$ (are thought to) belong to the same cluster, and 0 otherwise. This local version is interesting, as it allows to satisfy the variance condition with a smaller σ, which improves the MT-OMD regret bound. Note that the proof of Theorem 1 can be readily adapted to this definition by considering the class of interaction matrices $\{A(b) = I_N + bL^W\}$. The bound however features $\max_{i \leq N}[A(b)^{-1}]_{ii}$, which depends on W in a nontrivial way and requires a case by case analysis, preventing from stating a general result for an arbitrary W. Considering even more general matrices A, i.e., that do not write as $I_N + bL$, suffers from the same problem (one then also needs to compute $B_{\boldsymbol{\psi}}(\boldsymbol{A}^{1/2}\boldsymbol{u}, \boldsymbol{A}^{1/2}\boldsymbol{v})$ on a case by case basis), and does not enjoy anymore the variance interpretation seen above. Furthermore, note that Proposition 2 is obtained by minimizing Equation (11) with respect to A. For matrices of the form $A(b)$, this tradeoff only depends on b, and is thus much easier to solve than for general matrices. Finally, we stress that local variances can be similarly considered on the simplex. Instead of involving the global $\boldsymbol{u}_j^{\max}$, the variance formula then features for each task/node a local maximum (respectively minimum) over its neighbours.

4 Algorithms

We now show that MT-OGD and MT-EG enjoy closed-form updates, making them easy to implement. Note that the MT-OGD derivation is valid for any matrix A positive definite, while MT-EG requires $A^{-1/2}$ to be stochastic. This is verified by matrices of the form $A = I_N + L^W$ (Lemma 12).

MT-OGD. Let $V = \{u \in \mathbb{R}^d \colon \|u\|_2 \leq D\}$, and $\sigma \leq 1$. Recall that $\boldsymbol{V} = V^{\otimes N} = \{\boldsymbol{u} \in \mathbb{R}^{Nd} \colon \|\boldsymbol{u}\|_{2,\infty} \leq D\}$, and $\boldsymbol{V}_{\|\cdot\|_2,\sigma} = \{\boldsymbol{u} \in \boldsymbol{V} \colon \mathrm{Var}_{\|\cdot\|_2}(\boldsymbol{u}) \leq \sigma^2\}$. Solving the first equation in Equation (8), we obtain that the iterate $\boldsymbol{x}_{t+1}$ produced by MT-OGD is the solution to

$$\min_{\boldsymbol{x} \in \mathbb{R}^{Nd}} \left\{\left\|\boldsymbol{x}_t - \eta_t \boldsymbol{A}^{-1}\boldsymbol{g}_t - \boldsymbol{x}\right\|_{\boldsymbol{A}}^2 \;:\; \|\boldsymbol{x}\|_{2,\infty} \leq D\right\}.$$

However, computing this update is made difficult by the discrepancy between the norms used in the objective and the constraint. A simple work around consists in considering the minimization over the Mahalanobis ball $\boldsymbol{V}_{\boldsymbol{A}} = \{\boldsymbol{u} \in \mathbb{R}^{Nd} \colon \|\boldsymbol{u}\|_{\boldsymbol{A}}^2 \leq (1 + b\sigma^2)ND^2\}$ instead. It is easy to check that $\boldsymbol{V}_{\|\cdot\|_2,\sigma} \subset \boldsymbol{V}_{\boldsymbol{A}}$, so that every result derived in Section 3 for MT-OGD remains valid (only the fact that comparators and iterates are in $\boldsymbol{V}_{\boldsymbol{A}}$ is actually used). With the substitution $\boldsymbol{y}_t = \boldsymbol{A}^{1/2}\boldsymbol{x}_t$ the MT-OGD update then rewrites (see Appendix B.1 for technical details)

$$\boldsymbol{y}_{t+1} = \mathrm{Proj}\Big(\boldsymbol{y}_t - \eta_t \boldsymbol{A}^{-1/2}\boldsymbol{g}_t, \sqrt{(1 + b\sigma^2)N}D\Big)\,, \tag{16}$$

where $\mathrm{Proj}(x, \tau) = \min\left\{1, \frac{\tau}{\|x\|_2}\right\}x$. Note that Equation (16) can be easily turned back into an update on $\boldsymbol{x}_t$ by making the inverse substitution. In practice however, $\boldsymbol{x}_t$ is only computed to make the predictions.

MT-EG. Using Equation (8) with $\boldsymbol{y}_t = \boldsymbol{A}^{1/2}\boldsymbol{x}_t$, MT-EG reads

$$\begin{aligned}
\tilde{\boldsymbol{y}}_{t+1} &= \operatorname*{arg\,min}_{\boldsymbol{y} \in \mathbb{R}^{Nd}} \; \langle \eta \boldsymbol{A}^{-1/2}\boldsymbol{g}_t, \boldsymbol{y}\rangle + B_{\boldsymbol{\psi}}(\boldsymbol{y}, \boldsymbol{y}_t)\,, \\
\boldsymbol{y}_{t+1} &= \operatorname*{arg\,min}_{\boldsymbol{y} \in \boldsymbol{A}^{1/2}(\boldsymbol{\Delta})} \; B_{\boldsymbol{\psi}}(\boldsymbol{y}, \tilde{\boldsymbol{y}}_{t+1})\,,
\end{aligned} \tag{17}$$

where ψ is the compound negative entropy regularizer such that $\psi(\boldsymbol{x}) = \sum_{i=1}^{N}\sum_{j=1}^{d} \boldsymbol{x}_j^{(i)} \ln\left(\boldsymbol{x}_j^{(i)}\right)$. One can show (see Appendix B.2 for details) that the update can be rewritten for all $i \leq N$ and $j \leq d$

$$\tilde{\boldsymbol{y}}_{t+1,j}^{(i)} = \boldsymbol{y}_{t,j}^{(i)} \ \exp\left(-\eta A_{ii_t}^{-1/2} g_{t,j} - 1\right) ,$$
$$\boldsymbol{y}_{t+1,j}^{(i)} = \frac{\tilde{\boldsymbol{y}}_{t+1,j}^{(i)}}{\sum_{k=1}^{d} \tilde{\boldsymbol{y}}_{t+1,k}^{(i)}} .$$

Combining both equations, we finally obtain

$$\boldsymbol{y}_{t+1,j}^{(i)} = \frac{\boldsymbol{y}_{t,j}^{(i)} \ e^{-\eta A_{ii_t}^{-1/2} g_{t,j}}}{\sum_{k=1}^{d} \boldsymbol{y}_{t,k}^{(i)} \ e^{-\eta A_{ii_t}^{-1/2} g_{t,k}}} . \tag{18}$$

Update Equation (18) enjoys a natural interpretation. Each block $\boldsymbol{y}^{(i)}$ is operating an individual standard EG update, but with gradient $A_{ii_t}^{-1/2} g_t$. When $A = I_N$, only the active block is updated. Otherwise, the update of block i is proportional to $A_{ii_t}^{-1/2}$, that quantifies the similarity between tasks i and i_t. Although this work only focuses on OMD for clarity, note that considering Follow-the-Regularized-Leader—see, e.g., (Orabona, 2019, Section 7)—with $\lambda = \psi(\boldsymbol{A}^{1/2}\cdot)$ would yield similar bounds. This would allow, for instance, the use of time-varying learning rates with entropic regularization.

5 Experiments

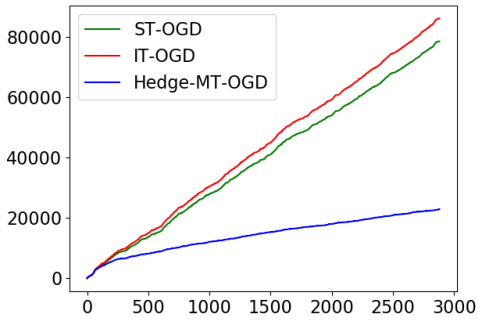
In this section, we empirically compare the performance of Hedge-MT-OGD/EG against two natural alternatives: an independent-task approach (IT-OGD/EG) where the agents do not communicate, and a single-task approach (ST-OGD/EG) where a single model is learned and shared by all agents. Note that both IT and ST approaches are special cases of MT-OMD, obtained respectively with the choices $b = 0$ (i.e., $\sigma^2 \geq 1$), or $b = +\infty$ (i.e., $\sigma^2 = 0$). In Appendix D we report an additional experiment where we empirically validate the dependence of the performance of MT-OGD on the task variance.

Online Gradient Descent. For this experiment, we use the *Lenk* dataset (Lenk et al., 1996; Argyriou et al., 2007). It consists of 2880 computer ratings in the range $\{1, 2, \ldots, 10\}$, made by 180 individuals (the tasks) on the basis of 14 binary features. Each computer is rated on a discrete scale from 0 to 10, expressing the likelihood of an individual buying that computer. We run Hedge-MT-OGD using the clique interaction matrix $A = (1 + N)I_N - \mathbb{1}\mathbb{1}^\top$ and the square loss. For all algorithms, the value of η is set according to the optimal theoretical value, see Proposition 2. In Hedge-MT-OGD, the variance σ^2 is learned in a set of 5 experts uniformly spaced over $[0, 1]$. For simplicity, we use $D = 1$ and compute the resulting Lipschitz constant accordingly. Results are reported in Figure 1(a).

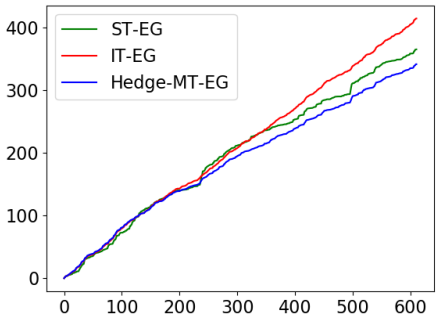
Exponentiated Gradient. For our second experiment, we consider *EMNIST*, a classification dataset consisting of 62 classes (images of digits, small and capital letters). To speed up computation, we reduced the number of features from 784 down to 10 through a standard dimensionality reduction method. We created 61 binary classification tasks by considering the 0 digit class against each other class. To each task, we assigned 10 examples (5 positive, 5 negative) randomly chosen from the set of examples for that task. We considered the linear logistic regression and ran Hedge-MT-EG with the parameterized clique interaction matrix $A(b) = (1 + b)I_N - b\mathbb{1}\mathbb{1}^\top/N$. The value of b is set according to the theoretical value (that depends on σ^2, see Proposition 8), while σ^2 is learned in a set of 5 experts uniformly spaced over $[0, 1]$. For all algorithms, the value of η is set according to the optimal theoretical values. Results are reported in Figure 1(b).

6 Conclusion

We introduced and analyzed MT-OMD, a multitask extension of OMD whose regret is shown to improve as the task variance, expressed in terms of the geometry induced by the regularizer, decreases. We provided a unifying analysis and a single algorithm that explains when is multitask acceleration possible based on the



(a) Lenk (OGD, cumulative square loss)



(b) EMNIST (EG, cumulative logistic loss)

Figure 1: Comparison between multitask (MT), independent-task (IT), and single-task (ST) OGD and EG on the *Lenk* and *EMNIST* datasets. We plot the cumulative losses against time. *Lenk* is known to work well in multi-task settings, and indeed Hedge-MT-OGD performs significantly better than both baselines. On the other hand, *EMNIST* has a variance significantly higher than *Lenk*. However, even in this unfavorable scenario, Hedge-MT-EG is still outperforming the baselines, though by a small margin.

current geometry, and how to achieve it. Natural and interesting directions for future research include: (1) analyzing the multitask acceleration in combination with other properties, such as strongly convex losses, and (2) designing and analyzing an extension of MT-OMD that is adaptive to the best interaction matrix.

References

Jacob Abernethy, Peter Bartlett, and Alexander Rakhlin. Multitask learning with expert advice. In *Proceedinds of the 20th International Conference on Computational Learning Theory*, pp. 484–498, 2007.

Gotz Alefeld and Norbert Schneider. On square roots of m-matrices. *Linear Algebra and its Applications*, 42: 119–132, 1982.

Pierre Alquier, The Tien Mai, and Massimiliano Pontil. Regret bounds for lifelong learning. In *Proceedinds of the 20th International Conference Artificial Intelligence and Statistics*, pp. 261–269, 2017.

Andreas Argyriou, Charles A Micchelli, Massimiliano Pontil, and Yiming Ying. A spectral regularization framework for multi-task structure learning. In *Proceedings of the 20th Annual Conference on Neural Information Processing Systems*, pp. 25–32, 2007.

Maria-Florina Balcan, Mikhail Khodak, and Ameet Talwalkar. Provable guarantees for gradient-based meta-learning. In *Proceedinds of the 36th International Conference on Machine Learning*, pp. 424–433, 2019.

Heinz H. Bauschke and Patrick L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.

Etienne Boursier, Mikhail Konobeev, and Nicolas Flammarion. Trace norm regularization for multi-task learning with scarce data. *arXiv preprint arXiv:2202.06742*, 2022.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

Giovanni Cavallanti, Nicolò Cesa-Bianchi, and Claudio Gentile. Linear algorithms for online multitask classification. *Journal of Machine Learning Research*, 11:2901–2934, 2010.

Nicolò Cesa-Bianchi, Claudio Gentile, and Giovanni Zappella. A gang of bandits. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems*, pp. 737–745, 2013.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

Ofer Dekel, Philip M Long, and Yoram Singer. Online learning of multiple tasks with a shared loss. *Journal of Machine Learning Research*, 8(10):2233–2264, 2007.

Giulia Denevi, Dimitris Stamos, Carlo Ciliberto, and Massimiliano Pontil. Online-within-online meta-learning. In *Proceedings of the 32th Annual Conference on Advances in Neural Information Processing Systems 32*, pp. 13089–13099, 2019.

Canh T Dinh, Tung T Vu, Nguyen H Tran, Minh N Dao, and Hongyu Zhang. Fedu: A unified framework for federated multi-task learning with Laplacian regularization. *arXiv preprint arXiv:2102.07148*, 2021.

Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 109–117, 2004.

Theodoros Evgeniou, Charles A Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6(4):615–637, 2005.

Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 1920–1930, 2019.

Claudio Gentile. The robustness of the p-norm algorithms. *Machine Learning*, 53(3):265–299, 2003.

Adam J Grove, Nick Littlestone, and Dale Schuurmans. General convergence results for linear discriminant updates. In *Proceedings 10th Annual Conference on Computational Learning Theory*, pp. 171–183, 1997.

Quanquan Gu, Zhenhui Li, and Jiawei Han. Learning a kernel for multi-task clustering. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, pp. 368–373, 2011.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4): 157–325, 2016.

Mark Herbster and Guy Lever. Predicting the labelling of a graph via minimum p-seminorm interpolation. In *Proceedings of the 22nd Conference on Learning Theory*, 2009.

Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *Journal of Machine Learning Research*, 13(1):1865–1890, 2012.

Pierre Laforgue, Andrea Della Vecchia, Nicolò Cesa-Bianchi, and Lorenzo Rosasco. Adatask: Adaptive multitask online learning. *arXiv preprint arXiv:2205.15802*, 2022.

Peter J Lenk, Wayne S DeSarbo, Paul E Green, and Martin R Young. Hierarchical bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*, 15(2):173–191, 1996.

Guangxia Li, Steven CH Hoi, Kuiyu Chang, Wenting Liu, and Ramesh Jain. Collaborative online multitask learning. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1866–1876, 2014.

Rui Li, Fenglong Ma, Wenjun Jiang, and Jing Gao. Online federated multitask learning. In *Proceedings of the 7th IEEE International Conference on Big Data*, pp. 215–220, 2019.

Keerthiram Murugesan, Hanxiao Liu, Jaime Carbonell, and Yiming Yang. Adaptive smoothed online multi-task learning. In *Proceedings of the 29th Annual Conference on Advances in Neural Information Processing Systems*, pp. 4296–4304, 2016.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Anastasia Pentina and Christoph H Lampert. Multi-task learning with labeled and unlabeled tasks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 2807–2816, 2017.

Gianluigi Pillonetto, Francesco Dinuzzo, and Giuseppe De Nicolao. Bayesian online multitask learning of gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(2):193–205, 2008.

Avishek Saha, Piyush Rai, Hal Daumé, Suresh Venkatasubramanian, et al. Online learning of multiple tasks and their relationships. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pp. 643–651, 2011.

Shai Shalev-Shwartz. Online learning: Theory, algorithms, and applications. *PhD thesis, The Hebrew University of Jerusalem, 2007.*, 2007.

Daniel Sheldon. Graphical multi-task learning. Technical report, Computing and Information Science Technical Reports, Cornell University, 2008.

Changjian Shui, Mahdieh Abbasi, Louis-Émile Robitaille, Boyu Wang, and Christian Gagné. A principled approach for learning task similarity in multitask learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 3446–3452, 2019.

Suvrit Sra. Fast projections onto mixed-norm balls with applications. *Data Mining and Knowledge Discovery*, 25(2):358–377, 2012.

Jonathan Tuck, Shane Barratt, and Stephen Boyd. A distributed method for fitting Laplacian regularized stratified models. *Journal of Machine Learning Research*, 22(60):1–37, 2021.

Yang Yang, Zhigang Ma, Yi Yang, Feiping Nie, and Heng Tao Shen. Multitask spectral clustering by exploring intertask correlation. *IEEE Transactions on Cybernetics*, 45(5):1083–1094, 2014.

Chi Zhang, Peilin Zhao, Shuji Hao, Yeng Chai Soh, Bu Sung Lee, Chunyan Miao, and Steven CH Hoi. Distributed multi-task classification: a decentralized online learning approach. *Machine Learning*, 107(4): 727–747, 2018.

Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

Yu Zhang and Dit Yan Yeung. A convex formulation for learning task relationships in multi-task learning. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pp. 733, 2010.