Logarithmic Regret for Online Gradient Descent Beyond Strong Convexity

Dan Garber
Technion - Israel Institute of Technology

Abstract

Hoffman's classical result gives a bound on the distance of a point from a convex and compact polytope in terms of the magnitude of violation of the constraints. Recently, several results showed that Hoffman's bound can be used to derive strongly-convex-like rates for first-order methods for *offline* convex optimization of curved, though not strongly convex, functions, over polyhedral sets. In this work, we use this classical result for the first time to obtain faster rates for *online convex optimization* over polyhedral sets with curved convex, though not strongly convex, loss functions. We show that under several reasonable assumptions on the data, the standard *Online Gradient Descent* algorithm guarantees logarithmic regret. To the best of our knowledge, the only previous algorithm to achieve logarithmic regret in the considered settings is the *Online Newton Step* algorithm which requires quadratic (in the dimension) memory and at least quadratic runtime per iteration, which greatly limits its applicability to large-scale problems. In particular, our results hold for *semi-adversarial* settings in which the data is a combination of an arbitrary (adversarial) sequence and a stochastic sequence, which might provide reasonable approximation for many real-world sequences, or under a natural assumption that the data is low-rank. We demonstrate via experiments that the regret of OGD is indeed comparable to that of ONS (and even far better) on curved though not strongly-convex losses.

1 Introduction

The celebrated *Online Gradient Descent* algorithm (OGD), originally due to [20], is a natural adaptation of the classical projected (sub)gradient descent algorithm for *offline* convex optimization, to the setting of *Online Convex Optimization* [5, 15]. The benefits of OGD are two folded: (i) in many problems of interest it performs very efficient iterations, which can often be executed in linear time (in the dimension), and (ii) it often guarantees optimal regret rates, mainly in terms of the length of the sequence T, e.g., $\sqrt{T}$ regret for arbitrary convex loss functions, and $\log T$ regret in case all loss functions are strongly-convex [6].

However, there exists a highly-important and wide class of loss functions, known as *exp-concave losses* [6], for which OGD does not guarantee optimal regret (in terms of T). For instance, the family of exp-concave losses capture important problems such as online linear regression with the square loss and online LASSO, online logistic regression, online portfolio selection, and more. While for exp-concave losses, OGD only guarantees regret that scales like $\sqrt{T}$, it is known that an online algorithm known as *Online Newton Step* (ONS), originally due to [6], guarantees $\log T$ regret (see also recent work [12] which gives an improved variant in terms of runtime and regret bound for low-rank data). On the downside, while OGD applies very efficient iterations in terms of runtime and memory requirements (e.g., when computing the gradient vector of the loss function and projecting onto the feasible set is computationally-cheap), ONS requires quadratic memory and to solve a linear system on each iteration (which requires at least quadratic runtime via efficient implementation). ONS also requires to compute a non-Euclidean projection on each iteration to enforce the constraints, which can be considerably more expensive than the Euclidean projection required by OGD (e.g., might require to use an iterative algorithm). Thus, despite the improved regret bound, ONS is often not applicable to large-scale problems. This naturally motivates the following question:

Can Online Gradient Descent be shown to enjoy a logarithmic regret bound for classes of loss functions beyond the class of strongly convex losses?

In this paper we take a step forward towards understanding the conditions under-which OGD can guarantee logarithmic regret, hence yielding both an efficient algorithm and an improved convergence rate in such settings. In particular, we focus on an important sub-class of the exp-concave losses: loss functions which can be written as a strongly-convex function applied to a linear transformation of the input variables. Such loss functions include important examples such as the square loss for linear regression, the online portfolio optimization loss, the logistic regression loss, and more. While such losses are not necessarily strongly convex in the entire space, they are strongly convex on a certain subspace, which corresponds to the row-span of the linear transformation. Our main result shows that when all loss functions are of this form, with linear transformations that satisfy certain consistency conditions, and the feasible set is a convex and compact polytope, the vanilla OGD algorithm, with a suitable choice of learning rate, indeed guarantees logarithmic regret. To the best of our knowledge, this is the first result to establish strongly-convex-like rates for OGD without strong convexity, on an important and wide class of applications.

Technically, at the heart of our result lies a classical result in convex analysis, originally due to Hoffman [8], which roughly speaking, bounds the distance of a point from a convex and compact polytope in terms of the magnitude of violation of the constraints describing the polytope. In case the feasible set is a polytope and the loss function is as described above (i.e., strongly convex applied to linear transformation), it can be shown that Hoffman's bound implies a property known as *quadratic growth*, which upper bounds the ℓ_2 distance between any feasible point and a feasible optimal solution, in terms of the distance in function values - a property well known to enable faster convergence rates in convex optimization settings (this is often also the main consequence of strong convexity needed in order to achieve fast rates for strongly-convex optimization).

Indeed, several recent works have used this classical result by Hoffman [8], to achieve fast rates without strong convexity for *offline optimization* problems, see for instance the recent works [14, 18, 9, 1, 19]. Importantly, all of these results consider only *stationary* settings, in which the objective function is fixed. As we show in the sequel, obtaining such fast rates results in the *online convex optimization* setting is considerably more challenging since, as opposed to strong convexity which is a property that holds in the entire space, and

hence, given a sequence of strongly convex functions, this property holds throughout the sequence, Hoffman's bound on the other-hand, is related to a specific subspace (which corresponds to the row-span of the linear transformation in the losses discussed above), and thus, given a sequence of such losses with different corresponding subspaces, these subspaces need not be, informally speaking, consistent with each other. Hence, a main contribution of this work is to formalize and analyze conditions under which this property could indeed be leveraged towards obtaining fast rates in a non-stationary online setting.

In particular, we show that our logarithmic-regret result holds for sequences which can be expressed as a combination of an arbitrary (adversarial) sequence and a stochastic sequence with certain stationary characteristics, which may potentially serve as reasonable approximation to many real-world data-streams, or when the data enjoys a low-rank structure. We report preliminary experimental results on both synthetic and real-world datasets which indeed show that OGD can outperform the Online Netwon Step method, both in terms of the regret and computational efficiency, on *non-strongly convex* sequences.

2 Preliminaries

Throughout this work we use $\|\cdot\|$ to denote the Euclidean norm for vectors and the spectral norm (i.e., largest singular value) for matrices. Also, for a compact set $\mathcal{P} \subset \mathbb{R}^d$ and a matrix $\mathbf{C} \in \mathbb{R}^{m \times d}$, we use the notation $\mathbf{C}\mathcal{P} := \{\mathbf{C}\mathbf{x} \mid \mathbf{x} \in \mathcal{P}\}$.

2.1 Convex optimization preliminaries

Definition 1. *Given a convex and compact set $\mathcal{K} \subset \mathbb{R}^d$ and a real-valued function f, differentiable over $\mathcal{K}$, we say f is G-Lipschitz over $\mathcal{K}$ if $\forall \mathbf{x} \in \mathcal{K} : \|\nabla f(\mathbf{x})\| \leq G$.*

In particular, if f is convex, differentiable and G-Lipschitz over a convex and compact $\mathcal{K}$, we have that $\forall \mathbf{x}, \mathbf{y} \in \mathcal{K}$:

$$f(\mathbf{x}) - f(\mathbf{y}) \leq (\mathbf{x} - \mathbf{y})^\top \nabla f(\mathbf{x}) \leq G\|\mathbf{x} - \mathbf{y}\|. \quad (1)$$

Definition 2. *Given a convex and compact set $\mathcal{K} \subset \mathbb{R}^d$ and a real-valued function f which is differentiable over $\mathcal{K}$, we say f is α-strongly convex over $\mathcal{K}$ if $\forall \mathbf{x}, \mathbf{y} \in \mathcal{K} : f(\mathbf{x}) \leq f(\mathbf{y}) + (\mathbf{x} - \mathbf{y})^\top \nabla f(\mathbf{x}) - \frac{\alpha}{2}\|\mathbf{x} - \mathbf{y}\|^2$.*

We recall the first-order optimality condition for convex differentiable functions (see for instance [2]): for any $\mathcal{K} \subset \mathbb{R}^d$, convex and compact, and a real-valued function f, convex and differentiable over $\mathcal{K}$,

we have that $\forall \mathbf{x} \in \mathcal{K}$, $\mathbf{x}^* \in \arg\min_{\mathbf{y} \in \mathcal{K}} f(\mathbf{y})$: $(\mathbf{x}^* - \mathbf{x})^\top \nabla f(\mathbf{x}^*) \leq 0$.

2.2 Online convex optimization preliminaries

We now briefly recall the setting of Online Convex Optimization (OCO). For a more in-depth introduction we refer the reader to [5, 15].

In the OCO problem, a decision maker (DM) is required to iteratively choose points in a fixed convex and compact set $\mathcal{K} \subset \mathbb{R}^d$. On each round t, after the DM makes his choice, i.e., chooses some $\mathbf{x}_t \in \mathcal{K}$, a convex function $f_t : \mathcal{K} \to \mathbb{R}$ is revealed, and the DM suffers the loss $f_t(\mathbf{x}_t)$. This process continues for T rounds, where T is assumed to be known in advanced. The goal is to design an algorithm for choosing the actions of the DM so to minimize a quantity called *regret*, which is given by

$$\text{regret}_T := \sum_{t=1}^{T} f_t(\mathbf{x}_t) - \min_{\mathbf{x} \in \mathcal{K}} \sum_{t=1}^{T} f_t(\mathbf{x}).$$

It is well known than an algorithm known as Online Gradient Descent, see Algorithm 1 below, can guarantee a $O(GD\sqrt{T})$ bound on the regret, where D is the ℓ_2 diameter of $\mathcal{K}$ and G is an ℓ_2 upper bound on the gradients of the functions $f_1, \ldots, f_T$ [20], which is in general optimal. It is also known that when all functions $f_1, \ldots, f_T$ are α-strongly convex, the same algorithm (though with different learning rate) guarantees $O((G^2/\alpha) \log T)$ regret [6], which is also optimal under this assumption [7].

Algorithm 1 Online (projected) Gradient Descent

1: $\mathbf{x}_1 \leftarrow$ some arbitrary point in $\mathcal{K}$
2: **for** $t = 1 \ldots T$ **do**
3:     $\mathbf{y}_{t+1} \leftarrow \mathbf{x}_t - \eta_t \nabla f_t(\mathbf{x}_t)$
4:     $\mathbf{x}_{t+1} \leftarrow \arg\min_{\mathbf{x} \in \mathcal{K}} \|\mathbf{x} - \mathbf{y}_{t+1}\|^2$
5: **end for**

2.3 Hoffman's bound and the quadratic growth property

Definition 3. *We say a matrix* $\mathbf{C} \in \mathbb{R}^{m \times d}$ *is* σ-*Hoffman with respect to a convex and compact polytope* $\mathcal{P} \subset \mathbb{R}^d$ *for some* $\sigma > 0$, *if for any vector* $\mathbf{c} \in \mathbb{R}^m$ *such that the set* $\mathcal{P}(\mathbf{C}, \mathbf{c}) := \{\mathbf{x} \in \mathcal{P} \mid \mathbf{C}\mathbf{x} = \mathbf{c}\}$ *is not empty, it holds that* $\forall \mathbf{x} \in \mathcal{P}$: $dist(\mathbf{x}, \mathcal{P}(\mathbf{C}, \mathbf{c}))^2 \leq \sigma^{-1} \|\mathbf{C}\mathbf{x} - \mathbf{c}\|^2$.

The following Lemma, originally due to Hoffman [8], shows that a Hoffman parameter bounded away from zero, always exists. Here we give the result in rephrased form. A proof is given in the appendix for completeness.

Lemma 1. *Let* $\mathcal{P} := \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$ *be a compact and convex polytope and let* $\mathbf{C} \in \mathbb{R}^{m \times d}$. *Given a vector* $\mathbf{c} \in \mathbb{R}^m$, *define the set* $\mathcal{P}(\mathbf{C}, \mathbf{c}) := \{\mathbf{x} \in \mathcal{P} \mid \mathbf{C}\mathbf{x} = \mathbf{c}\}$. *If* $\mathcal{P}(\mathbf{C}, \mathbf{c}) \neq \emptyset$, *then there exists* $\sigma > 0$ *such that* $\forall \mathbf{x} \in \mathcal{P}$: $dist(\mathbf{x}, \mathcal{P}(\mathbf{C}, \mathbf{c}))^2 \leq \sigma^{-1} \|\mathbf{C}\mathbf{x} - \mathbf{c}\|^2$. *Moreover, we have the bound* $\sigma \geq \min_{\mathbf{Q} \in \mathcal{M}} \lambda_{\min}(\mathbf{Q}\mathbf{Q}^\top)$, *where* $\mathcal{M}$ *is the set of all* $d \times d$ *matrices whose rows are linearly independent rows of the matrix* $\mathbf{M} := (\mathbf{A}^\top, \ \mathbf{C}^\top)^\top$, *and* $\lambda_{\min}(\cdot)$ *denotes the smallest non-zero eigenvalue.*

Definition 4. *. We say a function* $f : \mathbb{R}^d \to \mathbb{R}$ *has the quadratic growth property with parameter* κ *with respect to a compact and convex set* $\mathcal{K} \subset \mathbb{R}^d$, *if it holds that* $\forall \mathbf{x} \in \mathcal{K}$: $dist(\mathbf{x}, \mathcal{X}^*)^2 \leq \frac{2}{\kappa}(f(\mathbf{x}) - f^*)$, *where* $\mathcal{X}^* := \arg\min_{\mathbf{y} \in \mathcal{K}} f(\mathbf{y})$ *and* $f^* := \min_{\mathbf{y} \in \mathcal{K}} f(\mathbf{y})$.

The following lemma, which will be instrumental in the proof of our main result, demonstrates the connection between Hoffman's bound and the quadratic growth property for (possibly stochastic) convex objectives.

Lemma 2 (from Hoffman's bound to quadratic growth)**.** *Let* $\mathcal{P} \subset \mathbb{R}^d$ *be a convex and compact polytope. Let* $\mathcal{D}$ *be a distribution over pairs* $(g(\cdot), \mathbf{C}) \in (\mathbb{R}^m \to \mathbb{R}) \times \mathbb{R}^{m \times d}$ *satisfying:*

1. *for each pair* $(g, \mathbf{C})$ *in the support of* $\mathcal{D}$, *the function* g *is differentiable,* G-*Lipschitz, and* α-*strongly convex over* $\mathbf{C}\mathcal{P}$.

2. *the function* $F(\mathbf{x}) := \mathbb{E}_{(g, \mathbf{C}) \sim \mathcal{D}}[g(\mathbf{C}\mathbf{x})]$ *is differentiable over* $\mathcal{P}$.

3. *the expectation* $\mathbb{E}_{(g(\cdot), \mathbf{C}) \sim \mathcal{D}}[\mathbf{C}^\top \mathbf{C}]$ *exists.*

Let $\mathbf{C}_\mathcal{D} \in \mathbb{R}^{k \times d}$ *be such that* $\mathbf{C}_\mathcal{D}^\top \mathbf{C}_\mathcal{D} = \mathbb{E}_{(g(\cdot), \mathbf{C}) \sim \mathcal{D}}[\mathbf{C}^\top \mathbf{C}]$, *and denote by* σ *the Hoffman constant of* $\mathbf{C}_\mathcal{D}$ *w.r.t.* $\mathcal{P}$. *Finally, define* $\mathcal{X}^* := \arg\min_{\mathbf{y} \in \mathcal{P}} \{F(\mathbf{y}) := \mathbb{E}_{(g, \mathbf{C}) \sim \mathcal{D}}[g(\mathbf{C}\mathbf{y})]\}$. *Then, there exists* $\mathbf{c}_\mathcal{D} \in \mathbb{R}^k$ *such that* $\mathbf{x} \in \mathcal{X}^* \iff \mathbf{C}_\mathcal{D}\mathbf{x} = \mathbf{c}_\mathcal{D}$. *Moreover,* $\forall \mathbf{x} \in \mathcal{P}$: $dist(\mathbf{x}, \mathcal{X}^*)^2 \leq \frac{1}{\sigma}\|\mathbf{C}_\mathcal{D}\mathbf{x} - \mathbf{c}_\mathcal{D}\|^2 \leq \frac{2}{\alpha\sigma}(F(\mathbf{x}) - \min_{\mathbf{y} \in \mathcal{P}} F(\mathbf{y}))$.

Proof. For any $\mathbf{x} \in \mathcal{P}$ and $\mathbf{x}^* \in \mathcal{X}^*$ it holds that

$$
\begin{aligned}
F(\mathbf{x}^*) - F(\mathbf{x}) &= \mathbb{E}_{(g, \mathbf{C}) \sim \mathcal{D}}[g(\mathbf{C}\mathbf{x}^*) - g(\mathbf{C}\mathbf{x})] \\
&\underset{(a)}{\leq} \mathbb{E}_{(g, \mathbf{C}) \sim \mathcal{D}}\Big[(\mathbf{x}^* - \mathbf{x})^\top \mathbf{C}^\top \nabla g(\mathbf{C}\mathbf{x}^*) \\
&\qquad - \frac{\alpha}{2}\|\mathbf{C}(\mathbf{x}^* - \mathbf{x})\|^2\Big] \\
&= (\mathbf{x}^* - \mathbf{x})^\top \mathbb{E}_{(g, \mathbf{C}) \sim \mathcal{D}}\Big[\mathbf{C}^\top \nabla g(\mathbf{C}\mathbf{x}^*)\Big] \\
&\quad - \frac{\alpha}{2}(\mathbf{x}^* - \mathbf{x})^\top \mathbb{E}_{(g, \mathbf{C}) \sim \mathcal{D}}\Big[\mathbf{C}^\top \mathbf{C}\Big](\mathbf{x}^* - \mathbf{x}) \\
&\underset{(b)}{=} (\mathbf{x}^* - \mathbf{x})^\top \nabla F(\mathbf{x}^*) - \frac{\alpha}{2}\|\mathbf{C}_\mathcal{D}(\mathbf{x}^* - \mathbf{x})\|^2 \\
&\underset{(c)}{\leq} -\frac{\alpha}{2}\|\mathbf{C}_\mathcal{D}(\mathbf{x}^* - \mathbf{x})\|^2,
\end{aligned}
$$

where (a) follows since each $g(\cdot)$ in the support of $\mathcal{D}$ is differentiable and α-strongly convex over $\mathbf{C}\mathcal{P}$, (b) follows since $F(\mathbf{x})$ is differentaible over $\mathcal{P}$ and hence it's gradient vector is given by $\nabla F(\mathbf{x}) = \mathbb{E}_{(g,\mathbf{C})\sim\mathcal{D}}[\frac{d}{d\mathbf{x}}g(\mathbf{C}\mathbf{x})] = \mathbb{E}_{(g,\mathbf{C})\sim\mathcal{D}}[\mathbf{C}^\top\nabla g(\mathbf{C}\mathbf{x})]$, and (c) follows form the first-order optimality condition for $F(\cdot)$. Thus, $\forall \mathbf{x}^* \in \mathcal{X}^*, \mathbf{x} \in \mathcal{P}$:

$$\|\mathbf{C}_\mathcal{D}(\mathbf{x}^* - \mathbf{x})\|^2 \leq \frac{2}{\alpha}\left(F(\mathbf{x}) - F(\mathbf{x}^*)\right). \tag{2}$$

Thus, setting $\mathbf{c}_\mathcal{D} = \mathbf{C}_\mathcal{D}\mathbf{x}^*$ for some $\mathbf{x}^* \in \mathcal{X}^*$, directly gives the $\implies$ direction of the first part of the lemma

To prove the $\impliedby$ direction of the first part of the lemma, let $\mathbf{x} \in \mathcal{P}$ such that $\mathbf{C}_\mathcal{D}\mathbf{x} = \mathbf{c}_\mathcal{D}$. Then we have that

$$
\begin{aligned}
0 &= (\mathbf{x} - \mathbf{x}^*)\mathbf{C}_\mathcal{D}^\top\mathbf{C}_\mathcal{D}(\mathbf{x} - \mathbf{x}^*) \\
&= \mathbb{E}_{(g,\mathbf{C})\sim\mathcal{D}}[\|\mathbf{C}\mathbf{x} - \mathbf{C}\mathbf{x}^*\|^2] \\
&\geq \left(\mathbb{E}_{(g,\mathbf{C})\sim\mathcal{D}}[\|\mathbf{C}\mathbf{x} - \mathbf{C}\mathbf{x}^*\|]\right)^2.
\end{aligned}
$$

Since for each pair $(g,\mathbf{C})$ in the support of $\mathcal{D}$, g is convex and G-Lipschitz over $\mathbf{C}\mathcal{P}$, using Eq. (1) we have that

$$
\begin{aligned}
F(\mathbf{x}) - F(\mathbf{x}^*) &= \mathbb{E}_{(g,\mathbf{C})\sim\mathcal{D}}[g(\mathbf{C}\mathbf{x}) - g(\mathbf{C}\mathbf{x}^*)] \\
&\leq \mathbb{E}_{(g,\mathbf{C})\sim\mathcal{D}}[G\|\mathbf{C}\mathbf{x} - \mathbf{C}\mathbf{x}^*\|] = 0,
\end{aligned}
$$

meaning $\mathbf{x} \in \mathcal{X}^*$, which completes the proof of the first part of the lemma. The second part of the lemma follows directly form combining the first part of the lemma with Hoffman's bound (Lemma 1) and Eq. (2). $\square$

3 Informal Statement of Results and Examples

We now give an informal statement of our theoretical results, followed by several concrete examples to demonstrate possible applications. We then conclude the section by drawing a connection between our setting and *online exp-concave optimization* and the Online Newton Step algorithm.

3.1 Logarithmic regret for Online Gradient Descent without strong convexity

Suppose that the feasible set $\mathcal{P}$ is a convex and compact polytope in $\mathbb{R}^d$ and suppose all loss functions are of the form $f_t(\mathbf{x}) := g_t(\mathbf{C}_t\mathbf{x})$, where $g_t(\cdot)$ is differentiable and α_1-strongly convex. Suppose further, that there exists a matrix $\mathbf{M} \in \mathbb{R}^{k\times d}$ such that for all $t \in [T]$, $\mathbf{M}^\top\mathbf{M} \succeq \mathbb{E}[\mathbf{C}_t^\top\mathbf{C}_t] \succeq \alpha_2\mathbf{M}^\top\mathbf{M}$, where the expectation is with respect to possible randomness in the choice of $\mathbf{C}_t$. Then, we show there exists

a choice of step-sizes $\{\eta_t\}_{t\in[T]}$ such that OGD guarantees $O(\log T)$ regret (treating all other quantities as constants).

We note that while the requirement $\mathbf{M}^\top\mathbf{M} \succeq \mathbb{E}[\mathbf{C}_t^\top\mathbf{C}_t] \succeq \alpha_2\mathbf{M}^\top\mathbf{M}$ seems not standard at first glance, observe that when $\mathbf{C}_t$ is full-rank, and hence $f_t(\mathbf{x})$ is in particular strongly convex, this requirement holds trivially with $\mathbf{M} = \mathbf{I}$. Hence, this condition is natural for dealing with loss functions that are strongly convex only on a restricted subspace of $\mathbb{R}^d$, requiring them all to be consistent, at least in expectation, with the same subspace. In Subsection 3.2 we discuss several settings of interest in which it is reasonable to assume this requirement holds.

3.2 Examples of relevant settings and loss functions

Linear regression and Lasso: Consider the ℓ_p linear regression loss function $f_t(\mathbf{x}) := \frac{1}{2}\|\mathbf{A}_t\mathbf{x} - \mathbf{b}_t\|_p^2$, with $p \in (1,2]$. In particular, when $p = 2$ and the feasible polytope is an ℓ_1-ball, i.e., $\mathcal{P} := \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_1 \leq k\}$, for some $k > 0$, we get an online version of the famous LASSO problem [17].

For our log-regret result to hold for *deterministic data*, i.e., deterministic choice of $(\mathbf{A}_1, \mathbf{b}_1)\ldots(\mathbf{A}_t, \mathbf{b}_T)$, it must hold that row-span$(\mathbf{A}_1)$ = row-span$(\mathbf{A}_2)$ = $\cdots$ = row-span$(\mathbf{A}_T)$. This is reasonable if $\mathbf{A}_t \in \mathbb{R}^{m\times d}$ for a large enough value of m, and the data, i.e., the rows of $\mathbf{A}_1, \ldots, \mathbf{A}_T$ lie in a certain low-dimensional subspace [1].

A different *non-deterministic* setting of interest is a "semi-adversarial" model in which $\mathbf{A}_t := \tilde{\mathbf{A}}_t + \mathbf{N}_t$, where the matrices $\tilde{\mathbf{A}}_1, \ldots \tilde{\mathbf{A}}_T$ are arbitrary and $\mathbf{N}_t \sim \mathcal{D}$ i.i.d. for all $t \in [T]$, for some fixed (yet unknown) distribution $\mathcal{D}$. Then, a sufficient condition for our fast OGD rate to hold (in expectation), is that $\mathbb{E}_\mathcal{D}[\mathbf{N}] = \mathbf{0}$ and $\mathbb{E}_\mathcal{D}[\mathbf{N}^\top\mathbf{N}] \succeq \alpha\tilde{\mathbf{A}}_t^\top\tilde{\mathbf{A}}_t$ for all $t \in [T]$, for some $\alpha > 0$. That is, the data can be faithfully modeled as a deterministic sequence perturbed by a well-conditioned stochastic noise. For instance, such a model underlies the problem of *Universal Linear Filtering* studied in [13, 3].

Finally, if we can treat the data as generated by a stochastic mechanism that on each time t randomly samples $\mathbf{A}_t$ from a (unknown) distribution $\mathcal{D}_t$ (note we allow the distribution to change each round), then a sufficient condition for our log-regret result to hold is that there exists a matrix $\mathbf{M}$ and $\alpha > 0$ such that for all $t \in [T]$: $\mathbf{M}^\top\mathbf{M} \succeq \mathbb{E}_{\mathbf{A}_t\sim\mathcal{D}_t}[\mathbf{A}_t^\top\mathbf{A}_t] \succeq \alpha\mathbf{M}^\top\mathbf{M}$.

---

[1]this may be natural to assume for instance, if the data is the output of some dimension reduction technique such as the wildly used *principal component analysis* procedure

Logistic regression: In online logistic regression, the loss on m data points organized in a matrix $\mathbf{A}_t \in \mathbb{R}^{m \times d}$, can be written as $f_t(\mathbf{x}) := \sum_{i=1}^m \log\left(1 + \exp(\mathbf{A}_t^{(i)\top}\mathbf{x})\right)$, where $\mathbf{A}_t^{(i)}$ denotes the ith row of the matrix $\mathbf{A}_t$. Observe $f_t(\mathbf{x})$ can be rewritten as $f_t(\mathbf{x}) := g(\mathbf{A}_t\mathbf{x})$ with $g(\mathbf{y}) := \sum_{i=1}^m \log\left(1 + e^{\mathbf{y}_i}\right)$. It not difficult to verify that for bounded $\mathbf{y}$, $g(\mathbf{y})$ is indeed strongly convex, and hence this problem also falls into our setting. As in the linear regression case, if the feasible set is a polytope (e.g., standard selections are a ℓ_1 or ℓ_∞ ball), then same assumptions on the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_T$ will allow to apply our log-regret result.

Online portfolio selection: In the online portfolio selection problem [5], the loss of a *rebalancing* portfolio $\mathbf{x}$ (a point in the unit simplex) on m consecutive trading rounds is given by $f_t(\mathbf{x}) = -\sum_{i=1}^m \log\left(\mathbf{A}_t^{(i)\top}\mathbf{x}\right)$, where the rows of $\mathbf{A}_t \in \mathbb{R}^{m \times d}$, $\mathbf{A}_t > \mathbf{0}$ (entrywise) encodes the asset prices on each round. Similarly to the logistic regression example, we can write $f_t(\mathbf{x}) := g(\mathbf{A}_t\mathbf{x})$ with $g(\mathbf{y}) := -\sum_{i=1}^m \log(\mathbf{y}_i)$. Again, it is not hard to verify that if $\mathbf{A}_t \geq r$ (entry-wise) for some $r > 0$, then $g(\cdot)$ is indeed strongly convex over the transformed simplex $\mathbf{A}_t\Delta_d := \{\mathbf{A}_t\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^d, \mathbf{x} \geq \mathbf{0}, \sum_{i=1}^d \mathbf{x}_i = 1\}$. Again, our log-regret result holds under the same assumptions on the data $\mathbf{A}_1, \ldots, \mathbf{A}_T$, as above.

3.3 Connection with exp-concavity and the Online Newton Step algorithm

A real-valued function f, twice-differentiable over a compact set $\mathcal{K} \subset \mathbb{R}^d$, is σ exp-concave on $\mathcal{K}$ if and only if $\forall \mathbf{x} \in \mathcal{K}: \nabla^2 f(\mathbf{x}) \succeq \sigma \nabla f(\mathbf{x})\nabla f(\mathbf{x})^\top$ [5].

Note that in case $f(\mathbf{x}) := g(\mathbf{C}\mathbf{x})$, where g is α-strongly convex and twice-differentiable over $\mathbf{C}\mathcal{K}$, denoting $G = \sup_{\mathbf{x} \in \mathbf{C}\mathcal{K}} \|\nabla g(\mathbf{x})\|$, we have that $\forall \mathbf{x} \in \mathcal{K}$:

$$\nabla^2 f(\mathbf{x}) = \mathbf{C}^\top \nabla^2 g(\mathbf{C}\mathbf{x})\mathbf{C} \underset{(a)}{\succeq} \alpha\mathbf{C}^\top\mathbf{C} \succeq \frac{\alpha}{G^2}\|\nabla g(\mathbf{C}\mathbf{x})\|^2\mathbf{C}^\top\mathbf{C}$$

$$\underset{(b)}{\succeq} \frac{\alpha}{G^2}\mathbf{C}^\top \nabla g(\mathbf{C}\mathbf{x})\nabla g(\mathbf{C}\mathbf{x})^\top\mathbf{C} = \frac{\alpha}{G^2}\nabla f(\mathbf{x})\nabla f(\mathbf{x})^\top,$$

where (a) follows since $g(\cdot)$ is α-strongly convex, and (b) follows since for any vector $\mathbf{y}$ we have that $\mathbf{y}^\top\mathbf{C}^\top\nabla g(\mathbf{C}\mathbf{x})\nabla g(\mathbf{C}\mathbf{x})^\top\mathbf{C}\mathbf{y} = ((\mathbf{C}\mathbf{y})^\top\nabla g(\mathbf{C}\mathbf{x}))^2 \leq \|\nabla g(\mathbf{C}\mathbf{x})\|^2 \cdot \|\mathbf{C}\mathbf{y}\|^2 = \|\nabla g(\mathbf{C}\mathbf{x})\|^2 \cdot \mathbf{y}^\top\mathbf{C}^\top\mathbf{C}\mathbf{y}$. Hence, $f(\mathbf{x})$ is α/G^2-exp-concave over $\mathcal{K}$.

Thus, if all loss functions are as above, i.e., $f_t(\mathbf{x}) := \mathbf{g}_t(\mathbf{C}_t\mathbf{x})$, and we let G be a uniform upper bound on the ℓ_2 norm of the gradients of $g_t(\cdot)$ and C be a uniform upper bound on the spectral norm of the matrices $\mathbf{C}_t$, the Online Newton Step (ONS) algorithm [5], guarantees regret bound: $\text{regret}_T(ONS) =$

$O\left(\frac{G^2}{\alpha} + CGD\right)d\log T$, where D is the ℓ_2 diameter of $\mathcal{K}$.

4 Logarithmic Regret for OGD Without Strong Convexity

In this section we present and prove our main result - a logarithmic regret bound for Online Gradient Descent (Algorithm 1) without strong-convexity. As discussed, our result holds under certain conditions on the data which are captured in the following assumption. In the following Subsection 4.1 we discuss several concrete examples in which this assumption holds.

Assumption 1. *Given a convex and compact polytope $\mathcal{P} \subset \mathbb{R}^d$, a distribution $\mathcal{D}$ over pairs $(g, \mathbf{C}) \in (\mathbb{R}^m \to \mathbb{R}) \times \mathbb{R}^{m \times d}$, is said to satisfy Assumption 1 with parameters $(\mathbf{M}, \bar{G}, \alpha_1, \alpha_2) \in \mathbb{R}^{k \times d} \times \mathbb{R}_+^3$ w.r.t. $\mathcal{P}$, if it holds that*

1. *each function g, part of a pair $(g, \mathbf{C})$ in the support of $\mathcal{D}$, is differentiable, α_1-strongly convex over $\mathbf{C}\mathcal{P} := \{\mathbf{C}\mathbf{x} \mid \mathbf{x} \in \mathcal{P}\}$ and G-Lipschitz over $\mathbf{C}\mathcal{P}$, for some finite $G > 0$*

2. *the function $F(\mathbf{x}) := \mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}}[g(\mathbf{C}\mathbf{x})]$ is differentiable over $\mathcal{P}$*

3. *$\forall \mathbf{x} \in \mathcal{P}$: $\mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}}\left[\|\frac{d}{d\mathbf{x}}g(\mathbf{C}\mathbf{x})\|^2\right] = \mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}}\left[\|\mathbf{C}^\top\nabla g(\mathbf{C}\mathbf{x})\|^2\right] \leq \bar{G}^2$*

4. *the expectation $\mathbb{E}_{(g(\cdot),\mathbf{C}) \sim \mathcal{D}}[\mathbf{C}^\top\mathbf{C}]$ exists and satisfies $\mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}}[\mathbf{C}^\top\mathbf{C}] \succeq \alpha_2\mathbf{M}^\top\mathbf{M}$.*

We can now state our main theorem, Theorem 1. While the theorem holds under quite general conditions, we refer the reader again to Section 3.2 for discussion of concrete applications.

Theorem 1. *[OGD Master Theorem] Fix a convex and compact polytope $\mathcal{P} \subset \mathbb{R}^d$. Consider a sequence of T distributions $\mathcal{D}_1, \ldots \mathcal{D}_T$ over $(\mathbb{R}^m \to \mathbb{R}) \times \mathbb{R}^{m \times d}$ which satisfy Assumption 1 with parameters $(\bar{\mathbf{C}}, \bar{G}, \alpha_1, \alpha_2)$ w.r.t. $\mathcal{P}$, where $\bar{\mathbf{C}}$ is a matrix satisfying $\bar{\mathbf{C}}^\top\bar{\mathbf{C}} = \frac{1}{T}\sum_{t=1}^T \mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}_t}[\mathbf{C}^\top\mathbf{C}]$. Suppose further that $\bar{\mathbf{C}}$ is σ-Hoffman w.r.t. $\mathcal{P}$. Let $f_1(\mathbf{x}) \ldots f_T(\mathbf{x})$ be a sequence of loss functions such that $f_t(\mathbf{x}) = g_t(\mathbf{C}_t\mathbf{x})$, with $(g_t, \mathbf{C}_t) \sim \mathcal{D}_t$ independently of the functions $\{f_\tau\}_{\tau \in [T] \setminus \{t\}}$. Then, applying Algorithm 1 with stepsize $\eta_t = \frac{1}{\alpha_1\alpha_2\sigma t}$, w.r.t. the losses $f_1 \ldots f_T$ and the polytope $\mathcal{P}$, guarantees that*

$$\max_{\mathbf{x} \in \mathcal{P}} \mathbb{E}_{f_1 \sim \mathcal{D}_1 \ldots f_T \sim \mathcal{D}_T}\left[\sum_{t=1}^T f_t(\mathbf{x}_t) - \sum_{t=1}^T f_t(\mathbf{x})\right] \leq$$

$$\frac{\alpha_1\alpha_2\sigma D^2}{2} + \frac{\bar{G}^2}{2\alpha_1\alpha_2\sigma}(1 + \ln T).$$

Before we can prove the theorem, we need the following technical lemma which extends Lemma 2 from a single stochastic objective to a sequence of stochastic objectives, and hence plays a key role in our regret analysis .

Lemma 3 (from Hoffman's bound to quadratic growth of a sequence). *Let $\mathcal{D}_1, \ldots, \mathcal{D}_T$ be distributions over $(\mathbb{R}^m \to \mathbb{R}) \times \mathbb{R}^{m \times d}$ satisfying Assumption 1 with parameters $(\bar{\mathbf{C}}, \bar{G}, \alpha_1, \alpha_2) \in \mathbb{R}^{k \times d} \times \mathbb{R}_+^3$, where $\bar{\mathbf{C}}$ is a matrix satisfying $\bar{\mathbf{C}}^\top \bar{\mathbf{C}} = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}_t}[\mathbf{C}^\top \mathbf{C}_t]$. Consider the function $F(\mathbf{x}) := \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}_t}[g(\mathbf{C}\mathbf{x})]$, and define the set of feasible minimizers: $\mathcal{X}^* := \arg\min_{\mathbf{x} \in \mathcal{P}} F(\mathbf{x})$. Then, it holds that $\forall \mathbf{x}^*, \mathbf{y}^* \in \mathcal{X}^*, t \in [T]$: $\mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}_t}[g(\mathbf{C}\mathbf{x}^*)] = \mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}_t}[g(\mathbf{C}\mathbf{y}^*)]$. Moreover, letting σ denote the Hoffman constant of $\bar{\mathbf{C}}$ w.r.t. the polytope $\mathcal{P}$, we have that $\forall \mathbf{x} \in \mathcal{P}, \mathbf{x}^* \in \mathcal{X}^*$: $dist(\mathbf{x}, \mathcal{X}^*)^2 \le \sigma^{-1} \| \bar{\mathbf{C}}(\mathbf{x} - \mathbf{x}^*) \|^2$.*

Proof. Consider a distribution $\mathcal{D}$ over $(\mathbb{R}^m \to \mathbb{R}) \times \mathbb{R}^{m \times d}$, described by the following sampling procedure: pick $t \in [T]$ uniformly at random, and then sample $(g, \mathbf{C}) \sim \mathcal{D}_t$. Clearly, it holds that

$$
\begin{aligned}
\forall \mathbf{x} \in \mathcal{P} : \ F(\mathbf{x}) &= \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}_t}[g(\mathbf{C}\mathbf{x})] \\
&= \mathbb{E}_{t \sim \mathrm{Uni}[T]}[\mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}_t}[g(\mathbf{C}\mathbf{x})]] \\
&= \mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}}[g(\mathbf{C}\mathbf{x})].
\end{aligned}
$$

Thus, it follows that $\mathcal{X}^* = \arg\min_{\mathbf{x} \in \mathcal{P}} \mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}}[g(\mathbf{C}\mathbf{x})]$.

Note that since each distribution $\mathcal{D}_t$ satisfies Assumption 1, it also satisfies the assumptions of Lemma 2. It can be easily verified that as a consequence, the distribution $\mathcal{D}$ also satisfies the assumptions of Lemma 2, and thus there exists a matrix $\bar{\mathbf{C}}$, satisfying $\bar{\mathbf{C}}^\top \bar{\mathbf{C}} = \mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}}[\mathbf{C}^\top \mathbf{C}]$, such that $\forall \mathbf{x}^*, \mathbf{y}^* \in \mathcal{X}^*$:

$$
\| \bar{\mathbf{C}}(\mathbf{x}^* - \mathbf{y}^*) \|^2 = (\mathbf{x}^* - \mathbf{y}^*)^\top \bar{\mathbf{C}}^\top \bar{\mathbf{C}}(\mathbf{x}^* - \mathbf{y}^*) = 0.
$$

By the definitions of $\bar{\mathbf{C}}$ and the distribution $\mathcal{D}$, it holds that

$$
\begin{aligned}
\bar{\mathbf{C}}^\top \bar{\mathbf{C}} &= \mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}}[\mathbf{C}^\top \mathbf{C}] \\
&= \mathbb{E}_{t \sim \mathrm{Uni}[T]}\left[\mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}_t}[\mathbf{C}^\top \mathbf{C}] \right] = \frac{1}{T} \sum_{t=1}^{T} \bar{\mathbf{C}}_t^\top \bar{\mathbf{C}}_t,
\end{aligned}
$$

where we define $\bar{\mathbf{C}}_t$ to be a matrix satisfying $\bar{\mathbf{C}}_t^\top \bar{\mathbf{C}}_t = \mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}_t}[\mathbf{C}^\top \mathbf{C}]$.

Thus, we have that

$$
\begin{aligned}
\forall \mathbf{x}^*, \mathbf{y}^* \in \mathcal{X}^* : \ 0 &= (\mathbf{x}^* - \mathbf{y}^*)^\top \bar{\mathbf{C}}^\top \bar{\mathbf{C}}(\mathbf{x}^* - \mathbf{y}^*) \\
&= \frac{1}{T} \sum_{t=1}^{T} (\mathbf{x}^* - \mathbf{y}^*)^\top \bar{\mathbf{C}}_t^\top \bar{\mathbf{C}}_t(\mathbf{x}^* - \mathbf{y}^*),
\end{aligned}
$$

which, since each $\bar{\mathbf{C}}_t^\top \bar{\mathbf{C}}_t$ is positive semidefinite, implies that $\forall \mathbf{x}^*, \mathbf{y}^* \in \mathcal{X}^*, t \in [T]$:

$$
(\mathbf{x}^* - \mathbf{y}^*)^\top \bar{\mathbf{C}}_t^\top \bar{\mathbf{C}}_t(\mathbf{x}^* - \mathbf{y}^*) = 0. \tag{3}
$$

Thus, fixing some $\mathbf{x}^*, \mathbf{y}^* \in \mathcal{X}^*$ and $t \in [T]$, it holds that

$$
\begin{aligned}
&\mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}_t}[g(\mathbf{C}\mathbf{x}^*) - g(\mathbf{C}\mathbf{y}^*)] \underset{(a)}{\le} \\
&\mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}_t}[G\|\mathbf{C}(\mathbf{x}^* - \mathbf{y}^*)\|] = \\
&G\sqrt{\left(\mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}_t}[\|\mathbf{C}(\mathbf{x}^* - \mathbf{y}^*)\|]\right)^2} \le \\
&G\sqrt{\mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}_t}[\|\mathbf{C}(\mathbf{x}^* - \mathbf{y}^*)\|^2]} = \\
&G\sqrt{\mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}_t}[(\mathbf{x}^* - \mathbf{y}^*)^\top \mathbf{C}^\top \mathbf{C}(\mathbf{x}^* - \mathbf{y}^*)]} = \\
&G\sqrt{(\mathbf{x}^* - \mathbf{y}^*)^\top \mathbb{E}_{(g,\mathbf{C}) \sim \mathcal{D}_t}[\mathbf{C}^\top \mathbf{C}](\mathbf{x}^* - \mathbf{y}^*)} = \\
&G\sqrt{(\mathbf{x}^* - \mathbf{y}^*)^\top \bar{\mathbf{C}}_t^\top \bar{\mathbf{C}}_t(\mathbf{x}^* - \mathbf{y}^*)} \underset{(b)}{=} 0,
\end{aligned}
$$

where (a) follows via Eq. (1) since each $g(\cdot)$ in the support of $\mathcal{D}_t$ is convex and G-Lipschitz for some finite $G > 0$, and (b) follows from Eq. (3). Thus, the first part of the lemma follows.

The second part of the lemma is a straightforward consequence of Lemma 2, when applied to the distribution $\mathcal{D}$, defined above. $\qquad \square$

Proof of Theorem 1. Let us denote the set of of minimizers in hindsight: $\mathcal{X}^* = \arg\min_{\mathbf{x} \in \mathcal{P}} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{f_t \sim \mathcal{D}_t}[f_t(\mathbf{x})]$.

Given the sequence of points generated by Algorithm 1 $\{\mathbf{x}_t\}_{t=1}^{T}$, we define the sequence $\{\mathbf{x}_t^*\}_{t=1}^{T}$ as follows: $\forall t \ge 1$: $\mathbf{x}_t^* := \arg\min_{\mathbf{x} \in \mathcal{X}^*} \|\mathbf{x} - \mathbf{x}_t\|^2$, i.e., $\mathbf{x}_t^*$ is the projection of $\mathbf{x}_t$ onto the set of optimal plays in hindsight $\mathcal{X}^*$.

Let us fix some $\mathbf{x}^* \in \mathcal{X}^*$. By an application of Lemma 3, it holds that

$$
\begin{aligned}
\max_{\mathbf{x} \in \mathcal{P}} \mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x})\right] &= \mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}^*)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*)\right].
\end{aligned}
$$

Thus, to prove the theorem, it suffices to upper bound $R_T := \mathbb{E}_{f_1 \ldots f_T}\left[\sum_{t=1}^{T} f_t(\mathbf{x}_t) - \sum_{t=1}^{T} f_t(\mathbf{x}_t^*)\right]$.

Let us also define the sequence $\{r_t := f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*)\}_{t=1}^{T}$. Throughout the rest of the proof we write ∇_t as a short notation for $\nabla f_t(\mathbf{x}_t)$.

As standard in the analysis of Online Gradient Descent, for every $t \ge 1$, we have that

$$\|\mathbf{x}_{t+1} - \mathbf{x}_t^*\|^2 \le \|\mathbf{y}_{t+1} - \mathbf{x}_t^*\|^2 = \|\mathbf{x}_t - \eta_t \nabla_t - \mathbf{x}_t^*\|^2$$
$$= \|\mathbf{x}_t - \mathbf{x}_t^*\|^2 - 2\eta_t (\mathbf{x}_t - \mathbf{x}_t^*)^\top \nabla_t + \eta_t^2 \|\nabla_t\|^2,$$

where the first inequality holds since $\mathbf{x}_{t+1}$ is the orthogonal projection of $\mathbf{y}_{t+1}$ onto $\mathcal{P}$. Rearranging and recalling that $f_t(\mathbf{x}) = g_t(\mathbf{C}_t \mathbf{x})$ for some $g_t, \mathbf{C}_t$, we have that

$$(\mathbf{C}_t(\mathbf{x}_t - \mathbf{x}_t^*))^\top \nabla g_t(\mathbf{C}_t \mathbf{x}_t) = (\mathbf{x}_t - \mathbf{x}_t^*)^\top \nabla_t \le \frac{1}{2\eta_t} \|\mathbf{x}_t - \mathbf{x}_t^*\|^2$$
$$- \frac{1}{2\eta_t} \|\mathbf{x}_{t+1} - \mathbf{x}_t^*\|^2 + \frac{\eta_t}{2} \|\nabla_t\|^2.$$

Since $g_t(\cdot)$ is α_1-strongly convex, we have

$$r_t = g_t(\mathbf{C}_t \mathbf{x}_t) - g_t(\mathbf{C}_t \mathbf{x}_t^*) \le \frac{1}{2\eta_t} \|\mathbf{x}_t - \mathbf{x}_t^*\|^2 - \frac{1}{2\eta_t} \|\mathbf{x}_{t+1} - \mathbf{x}_t^*\|^2$$
$$- \frac{\alpha_1}{2} \|\mathbf{C}_t(\mathbf{x}_t - \mathbf{x}_t^*)\|^2 + \frac{\eta_t}{2} \|\nabla_t\|^2.$$

By the definition of the sequence $\{\mathbf{x}_t^*\}_{t=1}^T$, we have

$$r_t \le \frac{1}{2\eta_t} \mathrm{dist}(\mathbf{x}_t, \mathcal{X}^*)^2 - \frac{1}{2\eta_t} \mathrm{dist}(\mathbf{x}_{t+1}, \mathcal{X}^*)^2$$
$$- \frac{\alpha_1}{2} \|\mathbf{C}_t(\mathbf{x}_t - \mathbf{x}_t^*)\|^2 + \frac{\eta_t \|\nabla_t\|^2}{2}.$$

Summing over all T iterations, rearranging and taking expectation on both sides, we have that

$$R_T \le \mathbb{E}_{f_1 \dots f_T} \left[\frac{1}{2\eta_1} \mathrm{dist}(\mathbf{x}_1, \mathcal{X}^*)^2 + \frac{\bar{G}^2}{2} \sum_{t=1}^T \eta_t \right.$$
$$+ \frac{1}{2} \sum_{t=2}^T \left(\frac{1}{\eta_t} \mathrm{dist}(\mathbf{x}_t, \mathcal{X}^*)^2 - \frac{1}{\eta_{t-1}} \mathrm{dist}(\mathbf{x}_t, \mathcal{X}^*)^2 \right.$$
$$\left. \left. - \alpha_1 \|\mathbf{C}_t(\mathbf{x}_t - \mathbf{x}_t^*)\|^2 \right) \right]$$
$$\le \frac{1}{2} \mathbb{E}_{f_1 \dots f_t} \left[\sum_{t=2}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \mathrm{dist}(\mathbf{x}_t, \mathcal{X}^*)^2 \right.$$
$$\left. - \alpha_1 \|\mathbf{C}_t(\mathbf{x}_t - \mathbf{x}_t^*)\|^2 \right] + \frac{D^2}{2\eta_1} + \frac{\bar{G}^2}{2} \sum_{t=1}^T \eta_t, \quad (4)$$

where we have used the fact that $\forall t \in [T]$: $\mathbb{E}_{\mathcal{D}_t}[\|\nabla f_t(\mathbf{x}_t)\|^2] \le \bar{G}^2$.

Note that since for all t, $\mathbf{C}_t$ is independent of $\mathbf{x}_t, \mathbf{x}_t^*$, we have that

$$\mathbb{E}_{f_1 \dots f_t} \left[\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \mathrm{dist}(\mathbf{x}_t, \mathcal{X}^*)^2 - \alpha_1 \|\mathbf{C}_t(\mathbf{x}_t - \mathbf{x}_t^*)\|^2 \right] =$$

$$\mathbb{E}_{f_1 \dots f_{t-1}} \left[\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \mathrm{dist}(\mathbf{x}_t, \mathcal{X}^*)^2 \right.$$
$$\left. - \alpha_1 (\mathbf{x}_t - \mathbf{x}_t^*)^\top \mathbb{E}_{f_t}[\mathbf{C}_t^\top \mathbf{C}_t](\mathbf{x}_t - \mathbf{x}_t^*) \right] =$$

$$\mathbb{E}_{f_1 \dots f_{t-1}} \left[\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \mathrm{dist}(\mathbf{x}_t, \mathcal{X}^*)^2 - \alpha_1 \|\bar{\mathbf{C}}_t(\mathbf{x}_t - \mathbf{x}_t^*)\|^2 \right], \quad (5)$$

where we let $\bar{\mathbf{C}}_t$ be such that $\bar{\mathbf{C}}_t^\top \bar{\mathbf{C}}_t = \mathbb{E}_{(g, \mathbf{C}_t) \sim \mathcal{D}_t}[\mathbf{C}_t^\top \mathbf{C}_t]$.

By Lemma 3 and the assumption of the theorem, it holds that

$$\forall t \in [T]: \quad \mathrm{dist}(\mathbf{x}_t, \mathcal{X}^*)^2 \le \sigma^{-1} \|\bar{\mathbf{C}}(\mathbf{x}_t - \mathbf{x}_t^*)\|^2$$
$$\le (\alpha_2 \sigma)^{-1} \|\bar{\mathbf{C}}_t(\mathbf{x}_t - \mathbf{x}_t^*)\|^2.$$

Thus, since the step-size η_t is monotonically non-increasing with t, we have that $\forall t \ge 2$:

$$\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \mathrm{dist}(\mathbf{x}_t, \mathcal{X}^*)^2 - \alpha_1 \|\bar{\mathbf{C}}_t(\mathbf{x}_t - \mathbf{x}_t^*)\|^2$$
$$\le \left(\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \frac{1}{\alpha_2 \sigma} - \alpha_1 \right) \|\bar{\mathbf{C}}_t(\mathbf{x}_t - \mathbf{x}_t^*)\|^2. \quad (6)$$

Combining Eq. (4), (5), (6), and plugging-in our choice of step-size $\eta_t = \frac{1}{\alpha_1 \alpha_2 \sigma t}$, we conclude that

$$R_T \le \frac{D^2}{2\eta_1} + \frac{\bar{G}^2}{2} \sum_{t=1}^T \eta_t \le \frac{\alpha_1 \alpha_2 \sigma D^2}{2} + \frac{\bar{G}^2}{2\alpha_1 \alpha_2 \sigma}(1 + \ln T).$$

$\square$

4.1 Applications of Theorem 1

Deterministic data: In case $g_1, \dots g_T : \mathbb{R}^m \to \mathbb{R}$ are arbitrary α_1-strongly convex and differentiable functions over $\mathbb{R}^m$, then a sufficient condition on the matrices $\mathbf{C}_1, \dots, \mathbf{C}_T$ for applying the result of Theorem 1 is that there exists a positive constant α_2 such that

$$\forall (i, j) \in [T] \times [T]: \quad \mathbf{C}_i^\top \mathbf{C}_i \succeq \alpha_2 \mathbf{C}_j^\top \mathbf{C}_j, \quad (7)$$

or in a different formulation: row-span($\mathbf{C}_1$) = row-span($\mathbf{C}_2$) = $\cdots$ = row-span($\mathbf{C}_T$).

A simple application of Theorem 1 upper-bounds the regret by $\frac{\alpha_1 \alpha_2 \sigma D^2}{2} + \frac{G^2}{2\alpha_1 \alpha_2 \sigma}(1 + \ln T)$, where σ is the Hoffman constant of $\frac{1}{T} \sum_{t=1}^T \mathbf{C}_t^\top \mathbf{C}_t$ w.r.t. the polytope $\mathcal{P}$.

Semi-adversarial data: A way to circumvent the limitation of condition (7), is to consider slightly "easier" data. In particular if we let $\{g_t(\cdot)\}_{t \in [T]}$ be as in the deterministic case, but we assume that the matrices $\mathbf{C}_1, \dots, \mathbf{C}_T$ are perturbed realizations of some underlying deterministic sequence $\tilde{\mathbf{C}}_1, \dots, \tilde{\mathbf{C}}_T$. That is, we let $\tilde{\mathbf{C}}_t$ be arbitrary, but the observed matrix $\mathbf{C}_t$ is a perturbed version given by $\mathbf{C}_t = \tilde{\mathbf{C}}_t + \mathbf{N}_t$, where $\mathbf{N}_t \sim \mathcal{D}$, where $\mathcal{D}$ is a fixed unknown distribution. Then, the condition in (7) could be easily replaced by the requirement: $\forall t \in [T]$, $\mathbb{E}_{\mathcal{D}}[\mathbf{N}] = \mathbf{0}$ and $\mathbb{E}_{\mathcal{D}}[\mathbf{N}^\top \mathbf{N}] \succeq \alpha_2 \tilde{\mathbf{C}}_t^\top \tilde{\mathbf{C}}_t$, where α_2, as before, is a positive constant.
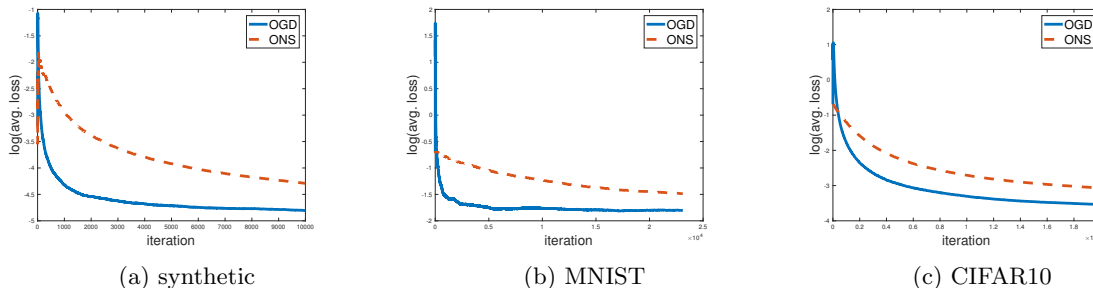
Figure 1: Comparing the (log) average loss vs. number of iterations for OGD and ONS.

An application of Theorem 1 upper-bounds the expected regret by $\frac{\alpha_1\alpha_2\sigma D^2}{2} + \frac{\bar{G}^2}{2\alpha_1\alpha_2\sigma}(1+\ln T)$, where σ is the Hoffman constant of the expected matrix $\frac{1}{T}\sum_{t=1}^T \mathbf{C}_t^\top \mathbf{C}_t + \mathbb{E}_\mathcal{D}[\mathbf{N}^\top\mathbf{N}]$, w.r.t. the polytope $\mathcal{P}$. As mentioned in Section 3.2, such a setting underlies for instance the *Universal Linear Filtering* problem studied in [13, 3].

Shifting stochastic data: Assuming $(g_t(\cdot), \mathbf{C}_t)$ is sampled out of a distribution $\mathcal{D}_t$ (possibly changing from round to round), under assumptions on $g_t(\cdot)$ as above, a sufficient condition on the stochastic matrices $\mathbf{C}_t$ for applying the result of Theorem 1 is that there exists some $\alpha_2 > 0$ such that for all $t_1, t_2 \in [T]$: $\mathbb{E}_{\mathbf{C}_1\sim\mathcal{D}_{t_1}}[\mathbf{C}_1^\top\mathbf{C}_1] \succeq \alpha_2\mathbb{E}_{\mathbf{C}_2\sim\mathcal{D}_{t_2}}[\mathbf{C}_2^\top\mathbf{C}_2]$, in which case we get the same bound as in the "semi-adversarial" case with σ being the Hoffman constant of $\frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\mathbf{C}\sim\mathcal{D}_t}[\mathbf{C}^\top\mathbf{C}]$ w.r.t. $\mathcal{P}$.

5 Experiments

In this section we provide empirical evidence for the performance of Online Gradient Descent on curved, though not strongly convex, losses. Since the computational advantage of OGD over competing methods is clear, we focus on demonstrating convergence in terms of the average loss. We consider the LASSO optimization problem, i.e., the loss function on each round t is $f_t(\mathbf{x}) := \frac{1}{2}\|\mathbf{a}_t^\top\mathbf{x} - b_t\|^2$ and the feasible polytope is an ℓ_1 ball. In all experiments we compare OGD with Online Newton Step (ONS).

Synthetic data: We compare OGD and ONS in an online stochastic setting. We fix the dimension to $d = 100$ and generate a random PSD matrix $\mathbf{M} \in \mathbb{R}^{d\times d}$ with rank = 50 and with decaying eigenvalues given by $\lambda_i = 10\cdot 0.8^{i-1}$ for all $i \in [50]$. We set $\mathbf{a}_t = \mathbf{v}_t^\top\mathbf{M}$ for a random unit vector $\mathbf{v}_t$, and $b_t := \mathbf{a}_t^\top\mathbf{w}^* + 0.1n_t$, where $\mathbf{w}^*$ is a fixed sparse vector chosen at random, and $n_t \sim \mathcal{N}(0,1)$. The radius of the feasible ℓ_1 ball is set to $r = 10$. Since determining the Hoffman constant is difficult in general, we heuristically set σ according the eigenvalue of the covariance matrix $\mathbf{A} = \frac{1}{T}\sum_{t=1}^T \mathbf{a}_t\mathbf{a}_t^\top$

which corresponds to the numerical rank of $\mathbf{A}$. We use this choice in all of our experiments which seems to work well. and set the step-size accordingly to $\eta_t = \frac{1}{\sigma t}$ (note $g_t(\cdot)$ in our case is 1-strongly convex). ONS is implemented as suggested in [5]. For both methods we plot the (log) average loss vs. number of iterations (we use $T = 10000$).

MNIST data: Next we experiment with the MNIST handwritten digit recognition dataset [11]. Specifically, we use the training dataset, keeping only the data related to digits $3, 5$. We set b_t by assigning value 1 to instances corresponding to the digit 5 and -1 to those corresponding to 3. Finally, in order to increase the amount of data, we replicate the data and concatenate twice. We set the radius of the feasible ℓ_1 ball to $r = 5$.

CIFAR10 data: we use the CIFAR10 tiny image dataset [10] which contains 50000 32x32 images in RGB format. We convert the images to grayscale and keep only the data related to the classes "automobile" and "truck", assigning the first the label $b_t = -1$ and the second the label $b_t = 1$. Here we also replicate the data twice and set $r = 5$.

The results for all datasets are presented in Figure 1. It is clearly observable that in all three cases OGD is comparable to ONS in terms of regret and even far better.

6 Acknowledgments

References

[1] Amir Beck and Shimrit Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. *Math. Program.*, 164(1-2):1–27, 2017.

[2] Stephen Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

[3] Dan Garber and Elad Hazan. Adaptive universal linear filtering. *IEEE Trans. Signal Processing,* 61(7):1595–1604, 2013.

[4] Osman Güler. *Foundations of optimization,* volume 258. Springer Science & Business Media, 2010.

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

[6] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning,* 69(2-3):169–192, 2007.

[7] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research,* 15(1):2489–2512, 2014.

[8] Alan J Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards,* 49(4), 1952.

[9] Hamed Karimi, Julie Nutini, and Mark Schmidt. *Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition,* pages 795–811. Springer International Publishing, Cham, 2016.

[10] Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.

[11] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE,* 86(11):2278–2324, 1998.

[12] Haipeng Luo, Alekh Agarwal, Nicolò Cesa-Bianchi, and John Langford. Efficient second order online learning by sketching. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain,* pages 902–910, 2016.

[13] Taesup Moon and Tsachy Weissman. Universal fir mmse filtering. *IEEE Transactions on Signal Processing,* 57(3):1068–1083, 2009.

[14] Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming,* pages 1–39, 2016.

[15] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning,* 4(2):107–194, 2012.

[16] Maurice Sion. On general minimax theorems. *Pacific Journal of mathematics,* 8(1):171–176, 1958.

[17] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological),* pages 267–288, 1996.

[18] Po-Wei Wang and Chih-Jen Lin. Iteration complexity of feasible descent methods for convex optimization. *Journal of Machine Learning Research,* 15(1):1523–1548, 2014.

[19] Yi Xu, Qihang Lin, and Tianbao Yang. Stochastic convex optimization: Faster local growth implies faster global convergence. In *International Conference on Machine Learning,* pages 3821–3830, 2017.

[20] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA,* pages 928–936, 2003.

A Proof of Hoffman's Lemma

For the convenience of the reader we first restate the lemma.

Lemma 4. *Let* $\mathcal{P} := \{\mathbf{x} \in \mathbb{R}^d \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$ *be a compact and convex polytope and let* $\mathbf{C} \in \mathbb{R}^{m \times d}$. *Given a vector* $\mathbf{c} \in \mathbb{R}^m$, *define the set* $\mathcal{P}(\mathbf{C}, \mathbf{c}) := \{\mathbf{x} \in \mathcal{P} \mid \mathbf{C}\mathbf{x} = \mathbf{c}\}$. *If* $\mathcal{P}(\mathbf{C}, \mathbf{c}) \neq \emptyset$, *then there exists* $\sigma > 0$ *such that*

$$\forall \mathbf{x} \in \mathcal{P}: \qquad dist(\mathbf{x}, \mathcal{P}(\mathbf{C}, \mathbf{c}))^2 \leq \sigma^{-1} \|\mathbf{C}\mathbf{x} - \mathbf{c}\|^2.$$

Moreover, we have the bound $\sigma \geq \min_{\mathbf{Q} \in \mathcal{M}} \lambda_{\min}\left(\mathbf{Q}\mathbf{Q}^\top\right)$, *where* $\mathcal{M}$ *is the set of all* $d \times d$ *matrices whose rows are linearly independent rows of the matrix* $\mathbf{M} := (\mathbf{A}^\top, \mathbf{C}^\top)^\top$, *and* $\lambda_{\min}(\cdot)$ *denotes the smallest non-zero eigenvalue.*

Proof. The proof is based on the proof in [4] (pages 299-301), though some details are different.

Let us define the matrix and vector

$$\tilde{\mathbf{A}} := \begin{pmatrix} \mathbf{A} \\ \mathbf{C} \\ -\mathbf{C} \end{pmatrix}, \qquad \tilde{\mathbf{b}} := \begin{pmatrix} \mathbf{b} \\ \mathbf{c} \\ -\mathbf{c} \end{pmatrix}.$$

Note that the following equivalence holds trivially

$$\mathcal{P}(\mathbf{C}, \mathbf{c}) := \{\mathbf{x} \in \mathbb{R}^d \mid \tilde{\mathbf{A}}\mathbf{x} \leq \tilde{\mathbf{b}}\}.$$

We can now write,

$$\begin{aligned}
\text{dist}(\mathbf{x}, \mathcal{P}(\mathbf{C}, \mathbf{c})) &= \min_{\mathbf{y} \in \mathcal{P}(\mathbf{C}, \mathbf{c})} \|\mathbf{x} - \mathbf{y}\| \\
&= \min_{\tilde{\mathbf{A}}\mathbf{y} \leq \tilde{\mathbf{b}}} \max_{\|\mathbf{u}\| \leq 1} (\mathbf{x} - \mathbf{y})^\top \mathbf{u} \\
&\overset{(a)}{=} \max_{\|\mathbf{u}\| \leq 1} \min_{\tilde{\mathbf{A}}\mathbf{y} \leq \tilde{\mathbf{b}}} (\mathbf{x} - \mathbf{y})^\top \mathbf{u} \\
&= \max_{\|\mathbf{u}\| \leq 1} \min_{\tilde{\mathbf{A}}\mathbf{w} \geq \hat{\mathbf{b}}} \mathbf{w}^\top \mathbf{u} \qquad \{\mathbf{w} := \mathbf{x} - \mathbf{y}, \quad \hat{\mathbf{b}} := \tilde{\mathbf{A}}\mathbf{x} - \tilde{\mathbf{b}}\} \\
&\overset{(b)}{=} \max_{\|\mathbf{u}\| \leq 1} \max\{\mathbf{q}^\top \hat{\mathbf{b}} \mid \tilde{\mathbf{A}}^\top \mathbf{q} = \mathbf{u}, \mathbf{q} \geq 0\} \\
&= \max\{\mathbf{q}^\top \hat{\mathbf{b}} \mid \|\tilde{\mathbf{A}}^\top \mathbf{q}\| \leq 1, \mathbf{q} \geq 0\}. \tag{8}
\end{aligned}$$

where (a) follows from minimax duality (which holds since the polytope $\tilde{\mathbf{A}}\mathbf{y} \leq \tilde{\mathbf{b}}$, by construction, is compact), see [16], and (b) follows from linear programming duality.

Let $\mathbf{q}^*$ be an optimal solution to the RHS of Eq. (8), and let $\mathbf{u}^*$ be a corresponding vector such that $\tilde{\mathbf{A}}^\top \mathbf{q}^* = \mathbf{u}^*$.

We can now write

$$\text{dist}(\mathbf{x}, \mathcal{P}(\mathbf{C}, \mathbf{c})) = \max\{\mathbf{q}^\top \left(\tilde{\mathbf{A}}\mathbf{x} - \tilde{\mathbf{b}}\right) \mid \tilde{\mathbf{A}}^\top \mathbf{q} = \mathbf{u}^*, \mathbf{q} \geq 0\}. \tag{9}$$

Note that the RHS of Eq. (9) is again a linear program with the feasible polytope being $\hat{\mathcal{P}} := \{\mathbf{q} \mid \tilde{\mathbf{A}}^\top \mathbf{q} = \mathbf{u}^*, \mathbf{q} \geq 0\}$. Hence, an optimal solution to RHS of (9) $\hat{\mathbf{q}}^*$ is without loss of generality a vertex of $\hat{\mathcal{P}}$. Hence, the non-zero entries of $\hat{\mathbf{q}}^*$ correspond to a set of linearly independent columns of $\tilde{\mathbf{A}}^\top$. That is, there exists a matrix $\mathbf{M}^*$, whose rows are taken from $\tilde{\mathbf{A}}$ and are linearly independent (i.e., $\mathbf{M}^*\mathbf{M}^{*\top}$ is positive definite), such that

$$1 \geq \|\mathbf{u}^*\| = \|\tilde{\mathbf{A}}^\top \hat{\mathbf{q}}^*\| \geq \|\hat{\mathbf{q}}^*\| \cdot \sigma_{\min}(\mathbf{M}^*) = \|\hat{\mathbf{q}}^*\| \cdot \sqrt{\lambda_{\min}(\mathbf{M}^*\mathbf{M}^{*\top})}.$$

Thus, we have that

$$\|\hat{\mathbf{q}}^*\| \leq \frac{1}{\sqrt{\lambda_{\min}(\mathbf{M}^*\mathbf{M}^{*\top})}}. \tag{10}$$

Note that since by definition $\hat{\mathbf{q}}^* \geq 0$, using the Cauchy-Schwarz inequality we have that

$$\text{RHS of (9)} \leq \|\hat{\mathbf{q}}^*\| \cdot \|(\tilde{\mathbf{A}}\mathbf{x} - \tilde{\mathbf{b}})_+\| \leq \|\hat{\mathbf{q}}^*\| \cdot \|\mathbf{C}\mathbf{x} - \mathbf{c}\|, \tag{11}$$

were for any vector $\mathbf{w}$ we let $\mathbf{w}_+$ denote the vector that corresponds to the subset of non-negative entries of $\mathbf{w}$. Note that the last inequality holds since $\mathbf{x}$ is feasible with respect to the polytope $\mathcal{P}$ (hence $\mathbf{A}\mathbf{x} - \mathbf{b} \leq 0$), and since for each row i of $\mathbf{C}$, it trivially holds that either $\mathbf{C}_i\mathbf{x} - \mathbf{c}_i \leq 0$, or $-\mathbf{C}_i\mathbf{x} + \mathbf{c}_i < 0$.

Combining Eq. (10), (11), yields the lemma. $\qquad\square$