Fast Rates in Stochastic Online Convex Optimization by Exploiting the Curvature of Feasible Sets

https://arxiv.org/abs/2402.12868

Taira Tsuchiya and Shinji Ito

The University of Tokyo & RIKEN

December 13, 2024

Neural Information Processing Systems 37 (NeurIPS 2024)

Online Convex Optimization (OCO)

for $t = 1, 2, \dots, T$ **do**

- Learner selects x_t from convex body $K \subset \mathbb{R}^d$ (K: **feasible set**)
- Environment reveals **convex loss function** $f_t \colon K \to \mathbb{R}$ (often bounded & Lipschitz)
- Learner incurs loss $f_t(x_t)$ and observes $\nabla f_t(x_t)$ (or f_t)

Online Convex Optimization (OCO)

for $t = 1, 2, \ldots, T$ **do**

- Learner selects x_t from convex body $K \subset \mathbb{R}^d$ (K: **feasible set**)
- Environment reveals **convex loss function** $f_t \colon K \to \mathbb{R}$ (often bounded & Lipschitz)
- Learner incurs loss $f_t(x_t)$ and observes $\nabla f_t(x_t)$ (or f_t)

Learner's Goal: Minimize the **(pseudo-)regret** R_T

$$\mathrm{R}_T = \max_{x \in K} \mathbb{E}\left[\sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x)\right] .$$

The **optimal decision** $x_\star$ is defined as $x_\star \in \arg\min_{x \in K} \mathbb{E}\left[\sum_{t=1}^{T} f_t(x)\right]$.

Online Convex Optimization (OCO)

for $t = 1, 2, \ldots, T$ **do**

Learner selects x_t from convex body $K \subset \mathbb{R}^d$ (K: **feasible set**)

Environment reveals **convex loss function** $f_t \colon K \to \mathbb{R}$ (often bounded & Lipschitz)

Learner incurs loss $f_t(x_t)$ and observes $\nabla f_t(x_t)$ (or f_t)

Learner's Goal: Minimize the **(pseudo-)regret** R_T

$$\mathrm{R}_T = \max_{x \in K} \mathbb{E}\left[\sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x)\right].$$

The **optimal decision** $x_\star$ is defined as $x_\star \in \arg\min_{x \in K} \mathbb{E}\left[\sum_{t=1}^{T} f_t(x)\right]$.

When loss function f_t is a linear function, *i.e.*, $f_t(\cdot) = \langle g_t, \cdot \rangle$ for some $g_t \in \mathbb{R}^d$, this problem is called **online linear optimization (OLO)**.

Application

- Stochastic (convex) optimization (via online-to-batch conversion)
  e.g., Stochastic Gradient Descent, AdaGrad, . . .
- Non-convex non-smooth optimization

Application

- Stochastic (convex) optimization (via online-to-batch conversion)
  e.g., Stochastic Gradient Descent, AdaGrad, ...
- Non-convex non-smooth optimization
- Online linear classification
  e.g., logistic loss $f_t(x) = \ln(1 + \exp(-y_t \cdot \langle x, z_t \rangle))$ for data-label pair (z_t, y_t)
- Online linear regression
  e.g., squared loss $f_t(x) = (\langle x, z_t \rangle - y_t)^2$

Application

- Stochastic (convex) optimization (via online-to-batch conversion)
  e.g., Stochastic Gradient Descent, AdaGrad, ...
- Non-convex non-smooth optimization
- Online linear classification
  e.g., logistic loss $f_t(x) = \ln(1 + \exp(-y_t \cdot \langle x, z_t \rangle))$ for data-label pair (z_t, y_t)
- Online linear regression
  e.g., squared loss $f_t(x) = (\langle x, z_t \rangle - y_t)^2$
- Bandits (multi-armed bandits, linear bandits, MDPs, ...)
- Online portfolio
- Learning in games
- ...

Lower Bound and Fast Rates for Curved Losses

Online Gradient Descent (OGD), $x_{t+1} \leftarrow \Pi_K(x_t - \eta_t \nabla f_t(x_t))$, achieves $\mathrm{R}_T = O(\sqrt{T})$ for Lipschitz continuous f_t [4].
The $O(\sqrt{T})$ bound cannot be improved in general [1].

However, this lower bound can be circumvented when the loss functions are curved! [1]

Definition (strongly convex and exp-concave functions)

A function $f\colon K \to (-\infty, \infty]$ is α-**strongly convex** (w.r.t. a norm $\|\cdot\|$) if for all $x, y \in K$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2} \|x - y\|^2 .$$

A function $f\colon K \to (-\infty, \infty]$ is β-**exp-concave** if $\exp(-\beta f(x))$ is concave.

- OGD with $\eta_t = \Theta(1/t) \to \mathrm{R}_T = O(\frac{1}{\alpha} \ln T)$ for α-strongly convex losses
- Online Newton Step (ONS) $\to \mathrm{R}_T = O(\frac{d}{\beta} \ln T)$ regret β-exp-concave losses

Lower Bound and Fast Rates for Curved Losses

Online Gradient Descent (OGD), $x_{t+1} \leftarrow \Pi_K(x_t - \eta_t \nabla f_t(x_t))$, achieves $\mathrm{R}_T = O(\sqrt{T})$ for Lipschitz continuous f_t [4].
The $O(\sqrt{T})$ bound cannot be improved in general [1].

However, this lower bound can be circumvented when the loss functions are curved! [1]

Definition (strongly convex and exp-concave functions)

A function $f\colon K \to (-\infty, \infty]$ is α-**strongly convex** (w.r.t. a norm $\|\cdot\|$) if for all $x, y \in K$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|x - y\|^2 .$$

A function $f\colon K \to (-\infty, \infty]$ is β-**exp-concave** if $\exp(-\beta f(x))$ is concave.

- OGD with $\eta_t = \Theta(1/t) \to \mathrm{R}_T = O(\frac{1}{\alpha} \ln T)$ for α-strongly convex losses
- Online Newton Step (ONS) $\to \mathrm{R}_T = O(\frac{d}{\beta} \ln T)$ regret β-exp-concave losses

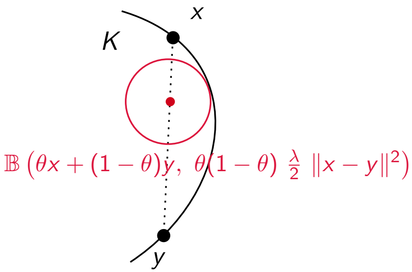
Q. Any other conditions under which we can circumvent the $\Omega(\sqrt{T})$ lower bound?

Exploiting the Curvature of Feasible Sets

Definition (strongly convex sets)

A convex body K is λ-**strongly convex** w.r.t. a norm $\|\cdot\|$ if

$$\forall x, y \in K, \forall \theta \in [0,1] \quad \theta x + (1-\theta)y + \theta(1-\theta)\frac{\lambda}{2}\|x-y\|^2 \cdot \mathbb{B}_{\|\cdot\|} \subseteq K \,.$$



Examples:

- ℓ_p-balls for $p \in (1,2]$
- Level set $\{x \colon f(x) \le r\}$ for a strongly convex and smooth function $f \colon \mathbb{R}^d \to \mathbb{R}$

Exploiting the Curvature of Feasible Sets

Definition (strongly convex sets)

A convex body K is λ-**strongly convex** w.r.t. a norm $\|\cdot\|$ if

$$\forall x, y \in K, \forall \theta \in [0,1] \quad \theta x + (1-\theta) y + \theta(1-\theta)\frac{\lambda}{2}\|x-y\|^2 \cdot \mathbb{B}_{\|\cdot\|} \subseteq K\,.$$

Theorem (Huang–Lattimore–György–Szepesvári, 2017 [2])

In online **linear** *optimization over λ-strongly convex sets,* **Follow-the-Leader (FTL)**, $x_t \in \arg\min_{x \in K} \sum_{s=1}^{t-1} \langle g_s, x \rangle$, *achieves (for G-Lipschitz losses)*

$$\mathrm{R}_T = O\left(\frac{G^2}{\lambda L} \ln T\right)$$

if there exists $L > 0$ such that $\|g_1 + \cdots + g_t\|_\star \geq tL$ for all $t \in [T]$ (growth condition).

This upper bound matches their lower bound.

Limitations of the Existing Approach

Limitations:

1. Only applicable to online **linear** optimization
   $\rightarrow$ Cannot leverage the curvature of loss functions
2. Can suffer a large regret when some ideal conditions (*e.g.*, the growth condition) are not satisfied
3. Curvature over the entire boundary of the feasible set is required

Limitations of the Existing Approach

Limitations:

1. Only applicable to online **linear** optimization
   $\rightarrow$ Cannot leverage the curvature of loss functions
2. Can suffer a large regret when some ideal conditions (*e.g.*, the growth condition) are not satisfied
3. Curvature over the entire boundary of the feasible set is required

Research Questions

1. Can we resolve these three limitations?
2. Are there any other characterizations of feasible sets for which we can achieve fast rates?

Sphere-enclosed Sets: A New Characterization of Feasible Sets

Definition (sphere-enclosed sets)

Let $K \subset \mathbb{R}^d$ be a convex body, $u \in \mathrm{bd}(K)$, and $f\colon K \to \mathbb{R}$. Then, convex body K is (ρ, u, f)-**sphere-enclosed** if there exists a ball $\mathbb{B}(c, \rho)$ with $c \in \mathbb{R}^d$ and $\rho > 0$ satisfying

1. $u \in \mathrm{bd}(\mathbb{B}(c, \rho))$
2. $K \subseteq \mathbb{B}(c, \rho)$
3. there exists $k > 0$ such that $u + k\nabla f(u) = c$

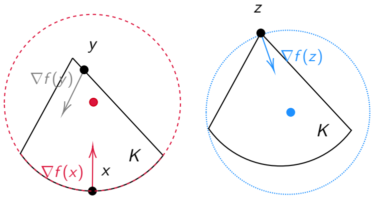


Figure: Examples of sphere-enclosed sets.

Main Result (1): Fast Rate over Sphere-enclosed Sets

Stochastic Environment: $f_1, f_2, \cdots \sim \mathcal{D}$, $f^\circ = \mathbb{E}_{f \sim \mathcal{D}}[f]$, and $x_\star = \arg\min_{x \in K} f^\circ(x)$
Adversarial Environment: $f_1, f_2, \ldots$ are fully adversarial

Main Result (1): Fast Rate over Sphere-enclosed Sets

Stochastic Environment: $f_1, f_2, \dots \sim \mathcal{D}$, $f^\circ = \mathbb{E}_{f\sim\mathcal{D}}[f]$, and $x_\star = \arg\min_{x\in K} f^\circ(x)$
Adversarial Environment: $f_1, f_2, \dots$ are fully adversarial

Theorem

Consider online **convex** *optimization. Suppose that K is $(\rho, x_\star, f^\circ)$-sphere-enclosed and that $\nabla f^\circ(x_\star) \neq 0$. Then, there exists an algorithm (***MetaGrad or universal online learning algorithm by van Erven–Koolen–van der Hoeven (2016, 2021)***) such that*

$$\mathrm{R}_T = O\left(\frac{G^2\rho}{\|\nabla f^\circ(x_\star)\|_2}\ln T\right) \quad \textit{in stochastic environments}$$

and $\mathrm{R}_T = O(GD\sqrt{T})$ *in adversarial environments. (D: diam of K, G: Lipschitzness of f_t)*

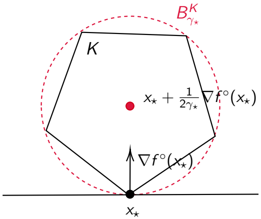
Main Result (1): Fast Rate over Sphere-enclosed Sets

Theorem

Consider online **convex** *optimization. Suppose that K is $(\rho, x_\star, f^\circ)$-sphere-enclosed and that $\nabla f^\circ(x_\star) \neq 0$. Then, there exists an algorithm (* **MetaGrad or universal online learning algorithm by van Erven–Koolen–van der Hoeven (2016, 2021)** *) such that*

$$\mathrm{R}_T = O\left(\frac{G^2\rho}{\|\nabla f^\circ(x_\star)\|_2} \ln T\right) \quad \textit{in stochastic environments}$$

and $\mathrm{R}_T = O(GD\sqrt{T})$ in adversarial environments. (D: diam of K, G: Lipschitzness of f_t)



Main Result (1): Fast Rate over Sphere-enclosed Sets

Theorem

Consider online **convex** *optimization. Suppose that K is $(\rho, x_\star, f^\circ)$-sphere-enclosed and that $\nabla f^\circ(x_\star) \neq 0$. Then, there exists an algorithm (* **MetaGrad or universal online learning algorithm by van Erven–Koolen–van der Hoeven (2016, 2021)** *) such that*

$$\mathrm{R}_T = O\left(\frac{G^2\rho}{\|\nabla f^\circ(x_\star)\|_2} \ln T\right) \quad \textit{in stochastic environments}$$

and $\mathrm{R}_T = O(GD\sqrt{T})$ *in adversarial environments. (D: diam of K, G: Lipschitzness of f_t)*

Matches the lower bound in Huang–Lattimore–György–Szepesvári (2017) [2]

Proof Overview (focusing only on T)

In stochastic environments, the regret is bounded from below by

$$\mathrm{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} (f^\circ(x_t) - f^\circ(x_\star))\right] \geq \mathbb{E}\left[\sum_{t=1}^{T} \langle \nabla f^\circ(x_\star), x_t - x_\star \rangle\right] \qquad \text{(convexity of } f^\circ\text{)}$$

$$\geq \mathbb{E}\left[\sum_{t=1}^{T} \gamma_\star \|x_t - x_\star\|_2^2\right] \quad \text{for some } \gamma_\star > 0 \qquad \text{(sphere-enclosedness of } K\text{)}$$

Proof Overview (focusing only on T)

In stochastic environments, the regret is bounded from below by

$$\mathrm{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} (f^\circ(x_t) - f^\circ(x_\star))\right] \geq \mathbb{E}\left[\sum_{t=1}^{T} \langle \nabla f^\circ(x_\star), x_t - x_\star \rangle\right] \qquad \text{(convexity of } f^\circ\text{)}$$

$$\geq \mathbb{E}\left[\sum_{t=1}^{T} \gamma_\star \|x_t - x_\star\|_2^2\right] \quad \text{for some } \gamma_\star > 0 \qquad \text{(sphere-enclosedness of } K\text{)}$$

There exists an algorithm achieving

$$\mathrm{R}_T \lesssim \mathbb{E}\left[\sqrt{\textstyle\sum_{t=1}^{T} \|x_t - x_\star\|_2^2 \ln T}\right].$$

Proof Overview (focusing only on T)

In stochastic environments, the regret is bounded from below by

$$\mathrm{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} (f^\circ(x_t) - f^\circ(x_\star))\right] \geq \mathbb{E}\left[\sum_{t=1}^{T} \langle \nabla f^\circ(x_\star), x_t - x_\star \rangle\right] \qquad \text{(convexity of } f^\circ\text{)}$$

$$\geq \mathbb{E}\left[\sum_{t=1}^{T} \gamma_\star \|x_t - x_\star\|_2^2\right] \quad \text{for some } \gamma_\star > 0 \qquad \text{(sphere-enclosedness of } K\text{)}$$

There exists an algorithm achieving

$$\mathrm{R}_T \lesssim \mathbb{E}\left[\sqrt{\textstyle\sum_{t=1}^{T} \|x_t - x_\star\|_2^2 \ln T}\right].$$

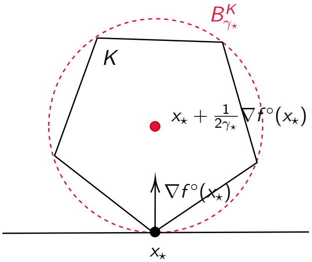
Combining upper and lower bounds of regret and Jensen's inequality gives

$$\mathrm{R}_T \lesssim \sqrt{\mathbb{E}\left[\sum_{t=1}^{T} \|x_t - x_\star\|_2^2\right] \ln T} - \gamma_\star \mathbb{E}\left[\sum_{t=1}^{T} \|x_t - x_\star\|_2^2\right] \underset{ax - bx^2 \leq a^2/(4b)}{\lesssim} \frac{\ln T}{\gamma_\star}. \quad \square$$

Check $\geq$

Consider a ball facing at $x_\star$:

$$B_\gamma^K = \mathbb{B}\left(x_\star + \frac{1}{2\gamma}\nabla f^\circ(x_\star)\,,\, \frac{1}{2\gamma}\|\nabla f^\circ(x_\star)\|_2\right) \subseteq \mathbb{R}^d$$



Check $\geq$

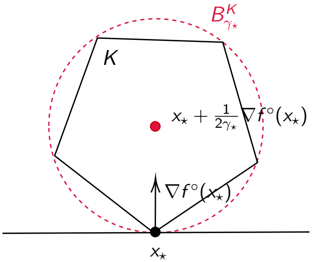
Consider a ball facing at $x_\star$:

$$B_\gamma^K = \mathbb{B}\left(x_\star + \frac{1}{2\gamma}\nabla f^\circ(x_\star)\,,\ \frac{1}{2\gamma}\|\nabla f^\circ(x_\star)\|_2\right) \subseteq \mathbb{R}^d$$

Observation

$z \in B_\gamma^K$ is equivalent to $\langle \nabla f^\circ(x_\star), z - x_\star\rangle \geq \gamma \|z - x_\star\|_2^2$.
Hence, from the $(\rho, x_\star, f^\circ)$-sphere-enclosedness of K, there exists γ so that $K \subseteq B_\gamma^K$, and thus

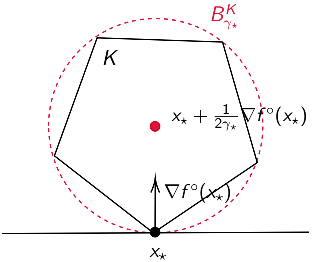
$$\langle \nabla f^\circ(x_\star), x_t - x_\star\rangle \geq \gamma \|x_t - x_\star\|_2^2\,.$$



Check $\geq$

Consider a ball facing at $x_\star$:

$$B_\gamma^K = \mathbb{B}\left(x_\star + \frac{1}{2\gamma}\nabla f^\circ(x_\star)\,,\ \frac{1}{2\gamma}\|\nabla f^\circ(x_\star)\|_2\right) \subseteq \mathbb{R}^d$$



Observation
$z \in B_\gamma^K$ is equivalent to $\langle \nabla f^\circ(x_\star), z - x_\star\rangle \geq \gamma\|z - x_\star\|_2^2$.
Hence, from the $(\rho, x_\star, f^\circ)$-sphere-enclosedness of K, there exists γ so that $K \subseteq B_\gamma^K$, and thus

$$\langle \nabla f^\circ(x_\star), x_t - x_\star\rangle \geq \gamma\|x_t - x_\star\|_2^2\,.$$

What is $\gamma_\star$?
One can set $\gamma_\star$ to $\gamma_\star = \sup\{\gamma \geq 0\colon K \subseteq B_\gamma^K\}$.
Since K is $(\rho, x_\star, f^\circ)$-sphere-enclosing, $\gamma_\star$ satisfies $\gamma_\star < \infty$ and $\frac{1}{2\gamma_\star}\|\nabla f^\circ(x_\star)\| = \rho$.

Benefits of Our Bound

Advantages against existing bounds:

1. Can achieve the $O(\ln T)$ regret if the boundary of K is curved around the optimal decision $x_\star$ or $x_\star$ in on corners
2. Can handle convex loss functions and thus the curvature of loss functions (*e.g.*, strong convexity or exp-concavity) can be simultaneously exploited
3. Can achieve $O(\sqrt{T})$ regret even in the worst-case scenarios

Limitations:

1. Achieve fast rates only in stochastic environments
   $\rightarrow$ Our regret bounds can be extended to corrupted stochastic environments! (omitted)

Q. Any other condition for which we can achieve fast rates?

Extending the Bound to Uniformly Convex Sets

Definition (uniformly convex sets)

A convex body K is (κ, q)-**uniformly convex w.r.t. a norm** $\|\cdot\|$ (or q-uniformly convex) if

$$\forall x, y \in K, \forall \theta \in [0,1] \quad \theta x + (1-\theta) y + \theta(1-\theta)\frac{\kappa}{2}\|x-y\|^{q} \cdot \mathbb{B}_{\|\cdot\|} \subseteq K\,.$$

Examples:

- ℓ_p-balls for $p \in (1, \infty)$
- $(\kappa, 2)$-uniformly convex set is κ-strongly convex

Extending the Bound to Uniformly Convex Sets

Definition (uniformly convex sets)

A convex body K is (κ, q)-**uniformly convex w.r.t. a norm** $\|\cdot\|$ (or q-uniformly convex) if

$$\forall x, y \in K, \forall \theta \in [0,1] \quad \theta x + (1-\theta) y + \theta(1-\theta)\frac{\kappa}{2}\|x-y\|^{q} \cdot \mathbb{B}_{\|\cdot\|} \subseteq K \,.$$

Theorem (Kerdreux–d'Aspremont–Pokutta, 2021 [3])

In online **linear** *optimization over* (κ, q)*-uniformly convex sets,* **Follow-the-Leader (FTL)**, $x_{t+1} \in \arg\min_{x \in K} \sum_{s=1}^{t-1} \langle g_s, x \rangle$, *achieves*

$$\mathrm{R}_T = O\left(\frac{G^{\frac{q}{q-1}}}{(\kappa L)^{\frac{1}{q-1}}} T^{\frac{q-2}{q-1}} \right)$$

if there exists $L > 0$ *such that* $\|g_1 + \cdots + g_t\|_\star \geq tL$ *for all* $t \in [T]$ *(growth condition).*

Extending the Bound to Uniformly Convex Sets

Theorem (Kerdreux–d'Aspremont–Pokutta, 2021 [3])

In online **linear** *optimization over* (κ, q)*-uniformly convex sets,* **Follow-the-Leader (FTL)**, $x_{t+1} \in \arg\min_{x \in K} \sum_{s=1}^{t-1} \langle g_s, x \rangle$, *achieves*

$$\mathrm{R}_T = O\left(\frac{G^{\frac{q}{q-1}}}{(\kappa L)^{\frac{1}{q-1}}} T^{\frac{q-2}{q-1}} \right)$$

if there exists $L > 0$ *such that* $\|g_1 + \cdots + g_t\|_\star \geq tL$ *for all* $t \in [T]$ *(growth condition).*

The bound $O\big(T^{\frac{q-2}{q-1}}\big)$ becomes smaller than $O(\sqrt{T})$ only when $q \in (2, 3)$.

Main Result (2): Faster Rates over Uniformly Convex Sets

Theorem

Consider online **convex** *optimization. Suppose that K is (κ, q)-uniformly convex and that $\nabla f^\circ(x_\star) \neq 0$. Then, there exists an algorithm such that*

$$\mathrm{R}_T = O\left(\frac{G^{\frac{q}{q-1}}}{(\kappa \|\nabla f^\circ(x_\star)\|_\star)^{\frac{1}{q-1}}} T^{\frac{q-2}{2(q-1)}} (\ln T)^{\frac{q}{2(q-1)}} \right) \quad \textit{in stochastic environments}$$

and $\mathrm{R}_T = O(GD\sqrt{T})$ in adversarial environments. (D: diam of K, G: Lipschitzness of f_t)

- Becomes $O(\ln T)$ when $q = 2$ and $\widetilde{O}(\sqrt{T})$ when $q \to \infty$, thus interpolating between the bound over the strongly convex sets and non-curved feasible sets
- Strictly better than the $O\left(T^{\frac{q-2}{q-1}}\right)$ bound in Kerdreux–d'Aspremont–Pokutta (2021) [3]

Summary

- Considered online convex optimization and introduced a new approach to achieve fast rates by exploiting the curvature of feasible sets
- Proved an $\mathrm{R}_T = O(\rho \ln T)$ regret bound for $(\rho, x_\star, f^\circ)$-sphere enclosed feasible sets
    1. Can exploit the curvature of loss functions
    2. Can achieve the $O(\ln T)$ regret bound only with local curvature properties
    3. Can work robustly even in environments where loss vectors do not satisfy the ideal conditions
- Proved the fast rates for uniformly convex feasible sets, which interpolates the $O(\ln T)$ regret over strongly convex sets and the $O(\sqrt{T})$ regret over non-curved sets

References

[1] Elad Hazan, Amit Agarwal, and Satyen Kale. "Logarithmic regret algorithms for online convex optimization". In: *Machine Learning* 69 (2007), pp. 169–192.

[2] Ruitong Huang et al. "Following the Leader and Fast Rates in Online Linear Prediction: Curved Constraint Sets and Other Regularities". In: *Journal of Machine Learning Research* 18.145 (2017), pp. 1–31.

[3] Thomas Kerdreux, Alexandre d'Aspremont, and Sebastian Pokutta. "Projection-Free Optimization on Uniformly Convex Sets". In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. Vol. 130. 2021, pp. 19–27.

[4] Martin Zinkevich. "Online convex programming and generalized infinitesimal gradient ascent". In: *Proceedings of the 20th International Conference on Machine Learning*. 2003, pp. 928–936.