Improved Dimensionality Dependence for Zeroth-Order Optimisation over Cross-Polytopes

Weijia Shao [1]

Abstract

This work proposes an algorithm improving the dimensionality dependence for gradient-free optimisation over cross-polytopes, which has many applications such as adversarial attacks, explainable AI and sparse regression. For bandit convex optimisation with two-point feedback over cross-polytopes, the state-of-the-art algorithms have a dimensionality dependence of $\mathcal{O}(\sqrt{d \log d})$, while the known lower bound is of the form $\Omega(\sqrt{d(\log d)^{-1}})$. We propose a mirror descent algorithm equipped with a symmetric version of the negative $\frac{1}{2}$-Tsallis entropy. Combined with an ℓ_1-ellipsoidal smoothing-based gradient estimator, the proposed algorithm guarantees a dimensionality dependence on $\mathcal{O}(\sqrt{d})$, which improves the state-of-the-art algorithms by a factor of $\sqrt{\log d}$. The idea can be further applied to optimising non-smooth and non-convex functions. We propose an algorithm with a convergence depending on $\mathcal{O}(d)$, which is the best-known dimensionality dependence.

1. Introduction

Gradient-free optimisation with two-point feedback (Agarwal et al., 2010) has attracted significant attention from the machine learning community. We are interested in the problem with decision set contained in the cross-polytopes, which is usually considered as a relaxation of the sparse constraint and has many applications, such as adversarial attack (Chen et al., 2018), explainable machine learning (Natesan Ramamurthy et al., 2020) and sparse cox regression (Liu et al., 2018). Decision sets contained in a cross-polytope define low-dimensional structures, making the problem less dimensionality-dependent. For the d-dimensional problem with convex and Lipschitz continuous objective functions, Duchi et al. have proved a lower bound of the iteration complexity depending on $\sqrt{d(\log d)^{-1}}$ and provided an algorithm with a convergence upper bound depending on $\sqrt{d \log d}$ if the objectives are smooth. For bandit convex optimisation with non-smooth loss functions, Shamir has proposed an algorithm with a regret upper bounded by $\mathcal{O}(\sqrt{Td} \log d)$, which has been improved to $\mathcal{O}(\sqrt{Td \log d})$ by Akhavan et al..

The algorithms described above are based on the idea of exploiting the geometry of the maximum normed space $(\mathbb{R}^d, \|\cdot\|_\infty)$. By combining gradient estimators with error bounded by $\mathcal{O}(d)$ and mirror descent (MD) with performance depending on $\mathcal{O}(\log d)$ (Duchi et al., 2015; Shamir, 2017; Akhavan et al., 2022), these algorithms can efficiently search the decision variables in the negative gradient direction in $(\mathbb{R}^d, \|\cdot\|_\infty)$. This idea has also been applied to optimising nonconvex but smooth objective functions in (Shao & Albayrak, 2023; Shao et al., 2022). It is impossible to further improve the gradient estimator (Kornowski & Shamir, 2023) or the searching strategy in $(\mathbb{R}^d, \|\cdot\|_\infty)$ (Orabona & Pál, 2015). Therefore, it is doubtful that the gap between the upper and lower bounds of the optimisation problem can be closed by following this idea.

This work is inspired by the Implicitly Normalised Forecaster (INF) (Audibert & Bubeck, 2010), which improves the dimensionality-dependence of the adversarial multi-armed bandit (MAB) problem from $\sqrt{d \log d}$ to $\sqrt{d}$. For each arm a in the action set and the probability p_a of choosing the arm, the importance-weighted estimator employed in INF is an unbiased estimator of the loss vector with a per-arm variance proportional to $\frac{1}{p_a}$, which has a strong dependence on the size of the action set. The exploitation strategy of INF, which is MD with negative $\frac{1}{2}$-Tsallis entropy (Audibert & Bubeck, 2010) and used to update p_a at each iteration, can normalise the variance to reduce the dimensionality dependence. Our idea is based on the similarity between the importance-weighted estimator and the ℓ_1-smoothing based two-point gradient estimator (Akhavan et al., 2022). By extending the randomisation over the ℓ_1-ball to the randomisation over the ℓ_1-ellipsoid, we construct

[1] Unit 2.6 Workplaces, Safety of Machinery, Operational Safety, Federal Institute for Occupational Safety and Health, Dresden, Germany. Correspondence to: Weijia Shao <Shao.Weijia@baua.bund.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

a gradient estimator with a pro-coordinate variance proportional to the inverse of the decision variable. To normalise the variance, the negative $\frac{1}{2}$-Tsallis entropy can not be directly applied since it works only for decision variables and gradients taken from the positive orthant. We propose a symmetric and strictly convex function retaining the curvature of the negative $\frac{1}{2}$-Tsallis entropy. MD equipped with the symmetric convex function can normalise the gradient estimator's variance, reducing the dimensionality dependence. Given a sequence of L_1-Lipschitz functions in $(\mathbb{R}^d, \|\cdot\|_1)$, our algorithm has a regret upper bounded by $\mathcal{O}(L_1\sqrt{Td})$, which improves the regret upper bound of state-of-the-art algorithms by a factor of $\sqrt{\log d}$.

Our idea can be further applied to optimising nonconvex and non-smooth objectives, which is computationally difficult (Zhang et al., 2020a) but bears considerable significance in the context of deep learning (Choromanski et al., 2018; Chen et al., 2017; Suh et al., 2022). Recent works focus on finding a (δ, ϵ)-stationary point defined in $(\mathbb{R}^d, \|\cdot\|_2)$ (Zhang et al., 2020a; Lin et al., 2022; Chen et al., 2023b;a; Kornowski & Shamir, 2023). Given fixed δ and ϵ, a vector x is considered to be stationary if there is a vector g in the convex hull of the generalised gradients (Clarke, 1990) at the points in a Euclidean ball centred at x, whose norm $\|g\|_2$ is less than ϵ. For a function that is L_2-Lipschitz continuous w.r.t $\|\cdot\|_2$, Lin et al. have introduced an algorithm finding a (δ, ϵ)-stationary point with $\mathcal{O}(d^{\frac{3}{2}}L_2^3\delta^{-1}\epsilon^{-4})$ function evaluations, which is improved to $\mathcal{O}(d^{\frac{3}{2}}L_2^2\delta^{-1}\epsilon^{-3})$ by the algorithm proposed by Chen et al., matching the optimal result w.r.t. δ and ϵ.

Since our motivation is related to deep learning, we are interested in the dimensionality dependence. Recently, Kornowski & Shamir have shown that the super-linear dependence on d can be improved and proposed an algorithm finding a stationary point within $\mathcal{O}(dL_2^2\delta^{-1}\epsilon^{-3})$ function evaluations. Note that the Lipschitz constant L_2 also contributes to the dimensionality dependence, which motivates us to consider a different geometry.

We focus on the objective functions that are L_1-Lipschitz w.r.t. $\|\cdot\|_1$. To define a meaningful stationary condition, we consider the neighbourhood defined by an ℓ_1-ball, but loosen the stationarity by requiring the maximum norm of the vector in the convex hull of the generalised gradient to be small. Combining our idea for bandit convex optimisation and the online-to-nonconvex conversion technique (Cutkosky et al., 2023), we construct an algorithm finding a (δ, ϵ)-stationary point (Lin et al., 2022) with $\mathcal{O}(dL_1^2\delta^{-1}\epsilon^{-3})$ noisy function evaluations, which is significantly better than the algorithms with dependence on $\mathcal{O}(d^{\frac{3}{2}}L_2^2)$. Compared to the algorithm with complexity $\mathcal{O}(dL_2^2\delta^{-1}\epsilon^{-3})$ introduced by Kornowski & Shamir, our algorithm has a similar dimensionality dependence if we consider the impact of equivalence of the norm on the stationarity $\|g\|_2 \leq \sqrt{d}\|g\|_\infty$, the neighbourhood $B_2(0, d^{-\frac{1}{2}}\delta) \subseteq B_1(0, \delta) \subseteq B_2(0, \delta)$ and the relation of the Lipschitz constant $L_1 \leq L_2 \leq \sqrt{d}L_1$. Our contributions and the discussion above are summarised in Table 1 and Table 2.

The rest of the paper is organised as follows. Section 2 introduces the notation and preliminary concepts. We present and analyse our algorithm for bandit convex optimisation in Section 3. In Section 4, we apply the idea introduced in Section 3 to stochastic optimisation of nonconvex and non-smooth objectives. Finally, we conclude our work in Section 5.

2. Preliminary

Throughout this paper, we consider optimisation problems in $\mathbb{R}^d$. We denote by $\|\cdot\|_p$ the p-norm for $p \geq 1$ and $\langle \cdot, \cdot \rangle$ the standard scalar product. We use $B_p(x, r)$ and $\partial B_p(x, r)$ for the closed ball and sphere in $(\mathbb{R}^d, \|\cdot\|_p)$ centred at x with radius r, respectively. Given a convex function ϕ, the Bregman divergence associated with ϕ is given by $\mathcal{B}_\phi(\cdot, \cdot)$.

For a convex function f, $\partial f(x)$ is the sub-differential of f at x and $\triangledown f(x)$ refers to any vector in $\partial f(x)$. If f is differentiable at x, we also use $\triangledown f(x)$ for the differential of f at x. For a nonconvex Lipschitz continuous but not everywhere differentiable function f, $\partial f(x)$ refers to the Clarke subgradient (Clarke, 1990) at $x \in \mathbb{R}^d$ given by

$$\partial f(x) = \mathrm{conv}\{\lim_{n\to\infty} \triangledown f(x_n) | x_n \to x\}.$$

For $\delta > 0$ and $p \geq 1$, we extend the definition of the Goldstein δ-subdifferential (Goldstein, 1977) of f by considering the neighbourhood defined by $B_p(0, \delta)$.

Definition 1. *Given $\delta > 0$, $p \geq 1$, the δ-subdifferential of a Lipschitz continuous function at x is defined by*

$$\partial_{\delta,p} f(x) := \mathrm{conv}(\cup_{y\in B_p(x,\delta)} \partial f(y)).$$

The (δ, ϵ)-stationary point is defined accordingly.

Definition 2. *A point x is a (δ, ϵ)-stationary point of a Lipschitz continuous function f if*

$$\|\triangledown f(x)\|_{\delta,p,q} := \min\{\|g\|_q | g \in \partial_{\delta,p} f(x)\} \leq \epsilon.$$

3. Bandit Convex Optimisation

We first consider the bandit convex optimisation over cross polytopes with two-point feedback, which can be considered as an iterative game between a player and an adversary. In each round t of the game, the player picks an action x_t from the decision set $\mathcal{K}$. Then, the adversarial selects a convex loss function f_t, incurring the loss $f_t(x_t)$. In the bandit setting, f_t is not revealed to the player, but the player is

Table 1. Bandit Convex Optimisation with Two-Point Feedback over Cross-polytopes

	Regret	Exploration	Exploitation
(Shamir, 2017)	$T^{\frac{1}{2}}d^{\frac{1}{2}}\log d$	ℓ_2-ball sampling	MD with Entropy
(Akhavan et al., 2022)	$T^{\frac{1}{2}}(d\log d)^{\frac{1}{2}}$	ℓ_1-ball sampling	MD with Entropy
Theorem 1	$T^{\frac{1}{2}}d^{\frac{1}{2}}$	ℓ_1-ellipsoid sampling	MD with symmetric Tsallis-Entropy
Lower Bound (Duchi et al., 2015)	$T^{\frac{1}{2}}d^{\frac{1}{2}}(\log d)^{-\frac{1}{2}}$		

Table 2. Gradient-Free Nonconvex and Non-smooth Optimisation

	neighbourhood	Stationarity	Smoothness	Complexity
(Lin et al., 2022)	ℓ_2-ball	$\|g\|_2 \leq \epsilon$	L_2-Lipschitz w.r.t. $\|\cdot\|_2$	$\mathcal{O}(d^{\frac{3}{2}}L_2^3\delta^{-1}\epsilon^{-4})$
(Chen et al., 2023a)	ℓ_2-ball	$\|g\|_2 \leq \epsilon$	L_2-Lipschitz w.r.t. $\|\cdot\|_2$	$\mathcal{O}(d^{\frac{3}{2}}L_2^2\delta^{-1}\epsilon^{-3})$
(Kornowski & Shamir, 2023)	ℓ_2-ball	$\|g\|_2 \leq \epsilon$	L_2-Lipschitz w.r.t. $\|\cdot\|_2$	$\mathcal{O}(dL_2^2\delta^{-1}\epsilon^{-3})$
Theorem 2	ℓ_1-ball	$\|g\|_\infty \leq \epsilon$	L_1-Lipschitz w.r.t. $\|\cdot\|_1$	$\mathcal{O}(dL_1^2\delta^{-1}\epsilon^{-3})$

allowed to query two points $f_t(x_t')$ and $f_t(x_t'')$ as feedback. The goal of the game is to control the cumulative regret of not choosing some action $x \in \mathcal{K}$

$$\sum_{t=1}^{T}(f_t(x_t) - f_t(x)).$$

In this section, we assume $\mathcal{K}$ is contained in a cross-polytope. The focus of our endeavour is to implement the mechanism of INF (Audibert & Bubeck, 2010) to operate seamlessly in this setting.

Exploration. We begin with constructing a gradient estimator using the ℓ_1-ellipsoidal smoothing, which is an extension of the ℓ_1-ball smoothing (Akhavan et al., 2022) and behaves similarly to the importance-weighted estimator in INF. Given Lipschitz continuous $f : \mathbb{R}^d \to \mathbb{R}$ and $x \in \mathbb{R}^d$, we define

$$g_{\gamma,\Lambda}(x) = \frac{d}{2\gamma}(f(x+\gamma\Lambda u) - f(x-\gamma\Lambda u))\Lambda^{-1}\operatorname{sgn}(u),$$

where $\gamma > 0$, u is sampled from the uniform distribution over $\partial B_1(0,1)$ and $\Lambda = \operatorname{diag}(\lambda)$ is a diagonal matrix with $\lambda_i > 0$ for $i = 1, \ldots, d$. Lemma 1 shows that $g_{\gamma,\Lambda}$ is an unbiased gradient estimator of smoothed f with variance related to $\|\lambda\|_2$.

Lemma 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L_1-Lipschitz w.r.t. $\|\cdot\|_1$ and $\Lambda = \operatorname{diag}(\lambda)$ some positive definite diagonal matrix. Define*

$$f_{\gamma,\Lambda} : \mathbb{R}^d \to \mathbb{R}, x \mapsto \mathbb{E}_\nu[f(x+\gamma\Lambda\nu)],$$

where ν is uniformly and randomly sampled from $B_1(0,1)$. Then, we have

a) $|f_{\gamma,\Lambda}(x) - f(x)| \leq \gamma\|\lambda\|_2 L_1$

b) $f_{\gamma,\Lambda}$ *is differentiable with* $\nabla f_{\gamma,\Lambda}(x) = \mathbb{E}[g_{\gamma,\Lambda}(x)]$

c) For u sampled from the uniform distribution over $\partial B_1(0,1)$, there is some constant c, such that

$$\operatorname{Var}_u(f(x+\gamma\Lambda u)) \leq \frac{c\gamma^2\|\lambda\|_2^2 L_1^2}{d^2}$$

holds for all $d \geq 3$.

Exploitation. The exploitation strategy of INF can be considered as an instance of MD equipped with a negative $\frac{1}{2}$-Tsallis entropy, which only takes values from the standard simplex. To obtain an algorithm for problems over the cross polytope with similar dimensionality dependence, we first extend the $\frac{1}{2}$-Tsallis entropy to the negative orthant. To achieve this, we define the following function.

$$\psi : \mathbb{R} \to \mathbb{R},$$
$$x \mapsto \begin{cases} \frac{2x}{\sqrt{\beta}} - 4\sqrt{\beta+x} + 4\sqrt{\beta}, & \text{if } x \geq 0 \\ -\frac{2x}{\sqrt{\beta}} - 4\sqrt{\beta-x} + 4\sqrt{\beta}. & \text{otherwise.} \end{cases} \quad (1)$$

For $\beta > 0$, the function is symmetric, twice continuously differentiable, strictly convex and locally self-concordant, which is proved in the next lemma.

Lemma 2. *ψ is twice continuously differentiable and strictly convex with the following properties.*

a) $\psi'(x) = \begin{cases} \frac{2}{\sqrt{\beta}} - \frac{2}{\sqrt{\beta+x}}, & \textit{if } x \geq 0 \\ -\frac{2}{\sqrt{\beta}} + \frac{2}{\sqrt{\beta-x}}. & \textit{otherwise} \end{cases}$

b) $\psi''(x) = (|x| + \beta)^{-\frac{3}{2}}$.

c) $(\psi')^{-1}(\theta) = \begin{cases} \frac{4\beta}{(2-\sqrt{\beta}\theta)^2} - \beta, & \textit{if } \frac{2}{\sqrt{\beta}} > \theta \geq 0 \\ -\frac{4\beta}{(2+\sqrt{\beta}\theta)^2} + \beta. & \textit{if } -\frac{2}{\sqrt{\beta}} < \theta \leq 0 \end{cases}$

d) For any $x, y \in \mathbb{R}$ satisfying

$$|\psi'(x) - \psi'(y)|\sqrt{|x| + \beta} \le 1,$$

we have $\mathcal{B}_\psi(y, x) \ge \frac{1}{16}(|x| + \beta)^{-\frac{3}{2}}(x - y)^2$.

We can construct a MD algorithm with an update-generating-function given by

$$\phi : \mathbb{R}^d \to \mathbb{R}, x \mapsto \sum_{i=1}^{d} \psi(x_i). \tag{2}$$

It follows from the last property in Lemma 2 that the update-generating function is self-concordant for large enough step-size. MD with the update-generating function defined in (2) has a "normalised" regret upper bound given controlled step size. We prove this in the next lemma.

Lemma 3. *Let $\mathcal{K} \subset \mathbb{R}^d$ be a closed convex set contained in a cross-polytope with radius D. Furthermore, let $\{\alpha_t\}$ be the sequence of non-decreasing stepsize and $\{g_t\}$ be any sequence of vectors in $\mathbb{R}^d$. Assume $\frac{|g_{t,i}|\sqrt{|x_{t,i}|+\beta}}{\alpha_t} \le 1$ for all t and i. Then, the update rule*

$$x_{t+1} = \arg\min_{x \in \mathcal{K}} \langle g_t, x\rangle + \alpha_t \mathcal{B}_\phi(x, x_t)$$

guarantees

$$\begin{aligned}&\sum_{t=1}^{T} \langle g_t, x_t - x\rangle \\ \le& 8\alpha_T D \beta^{-\frac{1}{2}} + \sum_{t=1}^{T} \frac{4}{\alpha_t} \sum_{i=1}^{d} (|x_{t,i}| + \beta)^{\frac{3}{2}} |g_{t,i}|^2.\end{aligned}$$

If x is taken from a cross-polytope with radius D and we pick $\lambda_{t,i} = \sqrt{|x_{t,i}| + \beta}$ for the gradient estimator, the variance term in the regret upper bound is normalised to

$$\sum_{i=1}^{d} (|x_{t,i}| + \beta)^{\frac{3}{2}} |g_{t,i}|^2 \le c(D + d\beta)^{\frac{3}{2}} d^{\frac{1}{2}} L_1^2,$$

which allows us to construct an algorithm with $\sqrt{d}$ dimensionality dependence. Algorithm 1 describes the obtained algorithm. In Theorem 1, we analyse the regret upper bound of the algorithm using constant and adaptive step sizes.

Theorem 1. *Let $\mathcal{K} \subseteq B_1(0, D)$ be a closed convex set and $\{f_t\}$ be a sequence of convex functions defined on $\mathcal{K}$. Assume $d \ge 3$ and L_1-Lipschitz continuity of $\{f_t\}$ w.r.t. $\|\cdot\|_1$. We run Algorithm 1 with update generating function (2). Then setting $\beta = \frac{D}{d}$, $\alpha_t = (2D)^{\frac{1}{2}} \max\{dL_1, \sqrt{T}L_1\}$ and $\gamma \le (Dd)^{\frac{1}{2}} T^{-\frac{1}{2}}$ guarantees*

$$\mathbb{E}[\sum_{t=1}^{T} (f_t(x_t) - f_t(x))] \le c_1 d^{\frac{1}{2}} D \max\{L_1\sqrt{T}, dL_1\},$$

Algorithm 1 Mirror Descent Framework

```
Input: Update Generating Function $\phi$
for $t = 1, \dots, T$ do
    $\lambda_{t,i} = \sqrt{|x_{t,i}| + \beta}$ for $i = 1, \dots, d$
    $g_t = \mathrm{GE}(f_t(\cdot), x_t, \lambda_t, \gamma)$
    Set stepsize $\alpha_t$
    $x_{t+1} = \arg\min_{x \in \mathcal{K}} \langle g_t, x\rangle + \alpha_t \mathcal{B}_\phi(x, x_t)$
end for
```

Algorithm 2 Gradient Estimator: $\mathrm{GE}(f, x, \lambda, \gamma)$

```
Sample $u$ uniformly from $\partial B_1(0, 1)$
Set $\Lambda = \mathrm{diag}(\lambda)$
Set $g = \frac{d}{2\gamma}(f(x + \gamma\Lambda u) - f(x - \gamma\Lambda u))\Lambda^{-1}\,\mathrm{sgn}(u)$
Return $g$
```

for some constant c_1 independent of D, L_1, d and T. Furthermore, setting

$$\alpha_t = (\frac{d^2}{4\gamma^2} \sum_{s=1}^{t} |f_s(x_s + \gamma\Lambda_s u_s) - f_s(x_s - \gamma\Lambda_s u_s)|^2)^{\frac{1}{2}}$$

ensures

$$\mathbb{E}[\sum_{t=1}^{T} (f_t(x_t) - f_t(x))] \le c_2 L_1 D \sqrt{dT},$$

for some constant c_2 independent of D L_1, d and T.

Proof. Denote by $\tilde{f}_t(\cdot) = \mathbb{E}_{v_t \sim B_1(0,1)} f_t(\cdot + \gamma\Lambda_t v_t)$ the smoothed f_t. $\tilde{f}_t$ is clear convex. Setting $\lambda_{t,i} = \sqrt{|x_{t,i}| + \beta}$ and $\beta = \frac{D}{d}$, we obtain

$$\begin{aligned}&\sum_{t=1}^{T} (f_t(x_t) - f_t(x)) \\ =& \sum_{t=1}^{T} (\tilde{f}_t(x_t) - \tilde{f}_t(x)) \\ &+ \sum_{t=1}^{T} (f_t(x_t) - \tilde{f}_t(x_t)) - \sum_{t=1}^{T} (f_t(x) - \tilde{f}_t(x)) \\ \le& \sum_{t=1}^{T} (\tilde{f}_t(x_t) - \tilde{f}_t(x)) + 2\sqrt{2D} L_1 \gamma T \\ \le& \sum_{t=1}^{T} \langle \nabla \tilde{f}_t(x_t), x_t - x\rangle + 2\sqrt{2D} L_1 \gamma T,\end{aligned}$$

where the first inequality uses the first property proved in Lemma 1 and $\|\lambda_t\|_2 = \sqrt{2D}$, the second inequality uses the convexity of $\tilde{f}_t$. Using the total law of expectation, we have

$$\sum_{t=1}^{T} \mathbb{E}[\mathbb{E}[\langle g_t - \nabla \tilde{f}_t(x_t), x_t - x\rangle | x_t]] = 0.$$

Combining the relations above, we have

$$\begin{aligned}
&\mathbb{E}[\sum_{t=1}^{T}(f_t(x_t) - f_t(x))] \\
\leq&\mathbb{E}[\sum_{t=1}^{T}\langle \nabla \tilde{f}_t(x_t), x_t - x\rangle] + 2\sqrt{2D}L_1\gamma T \\
=&\mathbb{E}[\sum_{t=1}^{T}\langle g_t, x_t - x\rangle] + 2\sqrt{2D}L_1\gamma T \qquad (3)\\
&+ \mathbb{E}[\sum_{t=1}^{T}\langle g_t - \nabla \tilde{f}_t(x_t), x_t - x\rangle] \\
=&\mathbb{E}[\sum_{t=1}^{T}\langle g_t, x_t - x\rangle] + 2\sqrt{2D}L_1\gamma T.
\end{aligned}$$

Next,

$$\begin{aligned}
|g_{t,i}| \leq& dL_1\|\Lambda_t u_t\|_1(|x_{t,i}| + \beta)^{-\frac{1}{2}} \\
\leq& dL_1\|\lambda_t\|_2\|u_t\|_2(|x_{t,i}| + \beta)^{-\frac{1}{2}} \\
\leq&(2D)^{\frac{1}{2}}dL_1(|x_{t,i}| + \beta)^{-\frac{1}{2}}
\end{aligned}$$

follows from the Lipschitz continuity and the fact $\|\lambda_t\|_2 \leq (2D)^{\frac{1}{2}}$. Setting $\alpha_1 =, \dots, = \alpha_t = (2D)^{\frac{1}{2}}\max\{dL_1, L_1\sqrt{T}\}$ ensures $\frac{|g_{t,i}|\sqrt{|x_{t,i}|+\beta}}{\alpha_t} \leq 1$, which allows us to apply Lemma 3 to obtain

$$\begin{aligned}
&\mathbb{E}[\sum_{t=1}^{T}\langle g_t, x_t - x\rangle] \\
\leq&8\sqrt{2}Dd^{\frac{1}{2}}\max\{dL_1, L_1\sqrt{T}\} \qquad (4)\\
&+ \frac{4}{(2D)^{\frac{1}{2}}L_1\sqrt{T}}\mathbb{E}[\sum_{t=1}^{T}\sum_{i=1}^{d}(|x_{t,i}| + \beta)^{\frac{3}{2}}|g_{t,i}|^2].
\end{aligned}$$

Define $l_t = f_t(x_t + \gamma\Lambda_t u_t)$ and $r_t = f_t(x_t - \gamma\Lambda_t u_t)$ Applying Lemma 1, we have

$$\begin{aligned}
&\mathbb{E}[\sum_{i=1}^{d}(|x_{t,i}| + \beta)^{\frac{3}{2}}|g_{t,i}|^2|x_t] \\
=&\frac{d^2}{4\gamma^2}\mathbb{E}[|l_t - r_t|^2|x_t]\sum_{i=1}^{d}(|x_{t,i}| + \beta)^{\frac{1}{2}} \\
\leq&\frac{d^2}{\gamma^2}\mathbb{E}[|l_t - \mathbb{E}_{u_t}[l_t]|^2|x_t]\sum_{i=1}^{d}(|x_{t,i}| + \beta)^{\frac{1}{2}} \qquad (5)\\
\leq&\frac{d^2}{\gamma^2}\operatorname{Var}(f_t(x_t + \gamma\Lambda_t u_t)|x_t)\sqrt{d}\sqrt{\sum_{i=1}^{d}(|x_{t,i}| + \beta)} \\
\leq&2\sqrt{2}D^{\frac{3}{2}}cd^{\frac{1}{2}}L_1^2
\end{aligned}$$

where the first inequality uses the symmetric distribution of u_t, the second inequality follows from the Cauchy-Schwarz inequality, and the last inequality uses the last property in Lemma 1 and the choice of λ_t. Combining with (4) and (5), we have

$$\begin{aligned}
&\mathbb{E}[\sum_{t=1}^{T}(f_t(x_t) - f_t(x))] \\
\leq&8\sqrt{2}Dd^{\frac{1}{2}}\max\{L_1\sqrt{T}, dL_1\} \qquad (6)\\
&+ 8cDd^{\frac{1}{2}}L_1\sqrt{T} + 2\sqrt{2D}L_1\gamma T.
\end{aligned}$$

We obtain the desired results by setting $\gamma \leq d^{\frac{1}{2}}D^{\frac{1}{2}}T^{-\frac{1}{2}}$. For the adaptive version, we have

$$\begin{aligned}
|g_{t,i}|\sqrt{|x_{t,i}| + \beta} =&\frac{d}{2\gamma}|l_t - r_t| \\
\leq&(\sum_{s=1}^{t}\frac{d^2}{4\gamma^2}|l_s - r_s|^2)^{\frac{1}{2}} \\
=&\alpha_t.
\end{aligned}$$

Thus, we can apply Lemma 3 and obtain

$$\begin{aligned}
&\sum_{t=1}^{T}\langle g_t, x_t - x\rangle \\
\leq&8D\alpha_T\beta^{-\frac{1}{2}} + \sum_{t=1}^{T}\frac{4}{\alpha_t}\sum_{i=1}^{d}(|x_{t,i}| + \beta)^{\frac{3}{2}}|g_{t,i}|^2 \\
=&8D\alpha_T\beta^{-\frac{1}{2}} + \sum_{t=1}^{T}\frac{4}{\alpha_t}\frac{d^2}{4\gamma^2}|l_t - r_t|^2\sum_{i=1}^{d}(|x_{t,i}| + \beta)^{\frac{1}{2}} \\
\leq&(8 + 8\sqrt{2})(Dd)^{\frac{1}{2}}\sqrt{\sum_{t=1}^{T}\frac{d^2}{4\gamma^2}|l_t - r_t|^2},
\end{aligned}$$

where the last inequality follows from Lemma 4 in (Orabona & Pál, 2018). Taking expectations of both sides and using Jensen's inequality, we obtain

$$\begin{aligned}
&\mathbb{E}[\sum_{t=1}^{T}\langle g_t, x_t - x\rangle] \\
\leq&\mathbb{E}[(8 + 8\sqrt{2})(Dd)^{\frac{1}{2}}\sqrt{\sum_{t=1}^{T}\frac{d^2}{4\gamma^2}|l_t - r_t|^2}] \\
\leq&(8 + 8\sqrt{2})(Dd)^{\frac{1}{2}}\sqrt{\mathbb{E}[\sum_{t=1}^{T}\frac{d^2}{4\gamma^2}|l_t - r_t|^2]} \\
\leq&(8 + 8\sqrt{2})(Dd)^{\frac{1}{2}}\sqrt{\sum_{t=1}^{T}\mathbb{E}[\mathbb{E}[\frac{d^2}{4\gamma^2}|l_t - r_t|^2|x_t]]} \\
\leq&(8 + 8\sqrt{2})(Dd)^{\frac{1}{2}}\sqrt{\sum_{t=1}^{T}\mathbb{E}[\mathbb{E}[\frac{d^2}{\gamma^2}|l_t - \mathbb{E}[l_t]|^2|x_t]]} \\
\leq&(8\sqrt{2} + 16)DL_1(cdT)^{\frac{1}{2}}.
\end{aligned}$$

Setting $\gamma \leq (Dd)^{\frac{1}{2}} T^{-\frac{1}{2}}$, we obtain the desired result. □

To get the variance normalised, we have to set the constant stepsize large enough so that Lemma 3 can be applied. With adaptive stepsizes, the condition ensuring self-concordance is always satisfied. Theorem 1 can be further generalised to the decision set contained in $B_p(0,1)$ and L_p-Lipschitz functions defined on $(\mathbb{R}^d, \|\cdot\|_p)$, for $p \in (1, \frac{\log d}{\log d - 1}]$. Akhavan et al. have discussed about the idea of applying the p-norm algorithm (Gentile, 2003) and ℓ_1-smoothing leading to a regret upper bound depending on $\mathcal{O}(\sqrt{d \log d})$. Our algorithm can be directly applied to the problem with a regret upper bound of the form $\mathcal{O}(L_p\sqrt{dT})$ because $B_p(0,1)$ is contained in $B_1(0, d^{1-\frac{1}{p}}) \subseteq B_1(0,e)$ and L_p-Lipschitz w.r.t. to $\|\cdot\|_p$ implies L_p-Lipschitz w.r.t. to $\|\cdot\|_1$.

It remains to show that the MD update in the last line of Algorithm 1 can be efficiently solved for $\mathcal{K} = B_1(0,1)$. Both constant and adaptive stepsize considered in Theorem 1 ensure $\frac{|g_{t,i}|}{\alpha_t} \leq \frac{1}{\sqrt{|x_{t,i}|+\beta}}$, i.e. $\alpha_t \nabla\phi(x_t) - g_t$ within the range of $\alpha_t \nabla\phi$. Thus, we can apply the two-step procedure

$$\tilde{x}_t = \nabla\phi^{-1}(\nabla\phi(x_t) - \frac{g_t}{\alpha_t})$$
$$x_{t+1} = \arg\min_{x \in B_1(0,1)} \phi(x) - \langle \nabla\phi(\tilde{x}_t), x\rangle.$$

From the convexity of ϕ, $x_{t+1} \in \partial B_1(0,1)$ must hold if $\tilde{x}_t \notin B_1(0,1)$. Thus, the problem is reduced to

$$x_{t+1} = \arg\min_{x \in \partial B_1(0,1)} \phi(x) - \langle \nabla\phi(\tilde{x}_t), x\rangle$$

for $\tilde{x}_t \notin B_1(0,1)$. W.l.o.g. we can assume $\tilde{x}_{t,i} \geq 0$. Using the Lagrange multipliers and setting the gradient of the auxiliary function to 0, we obtain that

$$\psi'(x_{t+1,i}) = \psi'(\tilde{x}_{t,i}) - \lambda + v_i$$

for some $\lambda \in \mathbb{R}$ and $v_i \geq 0$. $x_{t,i} \geq 0$ implies $x_{t+1,i} \geq 0$ and $\lambda > 0$. Let I be the index with $x_{t+1,i} \neq 0$. Using the complementary slackness, we have $x_{t+1,i} \neq 0$ for $v_i = 0$. Then, we must have

$$\sum_{i \in I} (\psi')^{-1}(\psi'(\tilde{x}_{t,i}) - \lambda) = 1,$$

which reduces the projection to the problem of sorting $\tilde{x}_{t,i}$ with a per iteration complexity $\mathcal{O}(d \log d)$.

4. Nonconvex Optimisation

The idea introduced in the previous section can be further applied to nonconvex and non-smooth stochastic optimisation to obtain an optimal dimensionality dependence. Formally, we wish to design a gradient-free algorithm for the stochastic optimisation problem of the form

$$\min_{x \in \mathcal{K}} \{F(x) := \mathbb{E}_\xi[f(x;\xi)]\}, \tag{7}$$

where F is a nonconvex and non-smooth function and ξ is a random variable. Following the setting of the previous work (Kornowski & Shamir, 2023), we assume the stochastic component $f(\cdot,\xi)$ is Lipschitz continuous with the Lipschitz constant depending on ξ.

Assumption 1. *For any ξ, $f(\cdot,\xi)$ is $L_1(\xi)$-Lipschitz w.r.t. $\|\cdot\|_1$. There is a $L_1 > 0$ such that $\mathbb{E}[L_1(\xi)^2] \leq L_1^2$.*

Without making additional assumptions on F, requiring the convergence of an optimisation algorithm towards a ϵ-Clarke stationary point is computationally intractable (Zhang et al., 2020b), and finding a point that is close enough to an ϵ-Clarke stationary point is similarly impossible (Kornowski & Shamir, 2022). Previous works (Cutkosky et al., 2023; Lin et al., 2022; Kornowski & Shamir, 2023) focus on finding a (δ,ϵ)-stationary point defined in the Euclidean space, i.e. a vector x satisfying $\|\nabla f(x)\|_{\delta,2,2} \leq \epsilon$. In case that f is L_2-Lipschitz w.r.t. $\|\cdot\|_2$, the δ-neighbourhood ensures $|f(x) - f(y)| \leq \delta L_2$ for all $y \in B_2(x,\delta)$ around x. Despite the possibility that none of the points in the ball are stationary, the objective values at these points are close. However, if we relax the continuity by assuming Lipschitzness w.r.t. $\|\cdot\|_1$, the difference between the objective values can be multiplied by a factor of $\sqrt{d}$ in the worst case. Therefore, we make the neighbourhood smaller but relax the stationarity by considering the condition $\|\nabla f(x)\|_{\delta,1,\infty} \leq \epsilon$. For the rest of the paper, we define

$$\partial_\delta f(x) = \operatorname{conv}(\cup_{y \in B_1(x,\delta)} \partial f(y)),$$

and

$$\|\nabla f(x)\|_\delta := \min\{\|g\|_\infty | g \in \partial_\delta f(x)\}.$$

Similar to the algorithm proposed in (Kornowski & Shamir, 2023), our idea is based on the relationship between the ℓ_1-ellipsoidal smoothing and the δ-subdifferential established by the next proposition.

Proposition 1. *Let f be L_1-Lipschitz w.r.t. $\|\cdot\|_1$ and $\Lambda = \operatorname{diag}(\lambda)$ for $\lambda \in \mathbb{R}^d_+$. Define*

$$f_{\gamma,\Lambda}(x) := \mathbb{E}_u f(x + \gamma\Lambda u)$$

for u randomly sampled from the uniform distribution over the unit ℓ_1-ball. Then we have $\nabla f_{\gamma,\Lambda}(x) \in \partial_{\gamma\|\lambda\|_2} f(x)$ and $\partial_\delta f_{\gamma,\Lambda}(x) \subseteq \partial_{\delta+\gamma\|\lambda\|_2} f(x)$ for all $x \in \mathbb{R}^d$.

Proposition 1 allows us to reduce the problem of finding a (δ,ϵ)-stationary point of F to the problem of finding a (δ,ϵ)-stationary point of $F_{\gamma,\Lambda}$, which is Lipschitz continuous and everywhere differentiable. Next, we apply the online-to-nonconvex technique (Cutkosky et al., 2023) to a sequence of smooth functions, which is described in Algorithm 3. The convergence of Algorithm 3 is analysed in Theorem 2.

Theorem 2. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a function satisfying Assumption 1 and*

$$\sup_{x \in \mathbb{R}^d} F(x_1) - F(x) \leq R.$$

Algorithm 3 Stochastic Nonsmooth Optimisation

```
Input: $\Delta > 0$, $K$ and $T$
Initialise: $x_1^1$, $y_1^1$
for $k = 1, \dots, K$ do
    for $t = 1, \dots, T$ do
        Sample $\xi_t^k$
        Sample $\eta_t^k$ uniformly from $[0, 1]$
        $z_t^k := y_t^k + \eta_t^k \Delta x_t^k$
        $\lambda_{t,i}^k := \sqrt{|x_{t,i}^k| + \frac{1}{d}}$ for $i = 1, \dots, d$
        $g_t^k := \mathrm{GE}(f(\cdot, \xi_t^k), z_t^k, \lambda_t^k, \gamma)$
        $l_t^k := f(z_t^k + \gamma \Lambda_t^k u_t^k, \xi_t^k)$
        $r_t^k := f(z_t^k - \gamma \Lambda_t^k u_t^k, \xi_t^k)$
        $\alpha_t^k := (\frac{d^2}{4\gamma^2} \sum_{s=1}^t |l_s^k - r_s^k|^2)^{\frac{1}{2}}$
        $x_{t+1}^k = \arg\min_{x \in B_1(0,1)} \langle g_t^k, x \rangle + \mathcal{B}_\phi(x, x_t^k)$
        $y_{t+1}^k := y_t^k + \Delta x_t^k$
    end for
    $\bar{z}^k := \frac{1}{T} \sum_{t=1}^T z_t^k$
    $y_1^{k+1} := y_{T+1}^k$
end for
Output: $\bar{z}$ uniformly sampled from $\{\bar{z}^1, \dots, \bar{z}^K\}$
```

Given $d \geq 3$, $\delta > 0$ and $\epsilon > 0$, there is a $N \in \mathcal{O}(\frac{RdL_1^2}{\epsilon^3 \delta})$, such that Algorithm 3 with $\Delta = \frac{\delta}{2T}$, $N = TK$,

$$T = \min\{\frac{N}{2}, (N\delta d^{\frac{1}{2}} L_1 R^{-1})^{\frac{2}{3}}\},$$

$$\gamma = \min\{\frac{\delta}{4}, T^{-1} N^{-\frac{1}{3}} \delta^{\frac{2}{3}} (RdL_1^2)^{\frac{1}{3}}\},$$

ϕ defined in (2) *and* GE *described in Algorithm 2 outputs a point $\bar{z}$ satisfying $\mathbb{E}[\|\nabla F(\bar{z})\|_\delta] \leq \epsilon$.*

Sketch of the proof. Define

$$\nabla_t^k := \mathbb{E}[\nabla F_{\gamma, \Lambda_t^k}(y_t^k + \eta_t^k \Delta x_t^k) | x_t^k].$$

It follows from Proposition 2 in (Cutkosky et al., 2023) that

$$F_{\gamma,\Lambda_t^k}(y) - F_{\gamma,\Lambda_t^k}(x) = \langle \mathbb{E}_\eta[\nabla F_{\gamma,\Lambda_t^k}(x + \eta(y - x))], y - x \rangle,$$

holds for η randomly sampled from the uniform distribution over $[0, 1]$ and all $x, y \in \mathbb{R}^d$. With this property, Algorithm 3 ensures the inequality described in Lemma 4

Lemma 4. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a function satisfying Assumption 1 and*

$$\sup_{x \in \mathbb{R}^d} F(x_1) - F(x) \leq R.$$

Then, running Algorithm 3 guarantees

$$\begin{aligned} -\sum_{k=1}^K \sum_{t=1}^T \langle g_t^k, v^k \rangle \leq & \sum_{k=1}^K \sum_{t=1}^T \langle g_t^k, x_t^k - v^k \rangle \\ & + \sum_{k=1}^K \sum_{t=1}^T \langle \nabla_t^k - g_t^k, x_t^k \rangle \\ & + \frac{R}{\Delta} + \frac{2\sqrt{2} N \gamma L_1}{\Delta}, \end{aligned}$$

for any $v^1, \dots, v^k$.

Setting $v^k = -\partial \|\cdot\|_\infty (\frac{1}{T} \sum_{t=1}^T \nabla_t^k)$, we decompose the convergence into

$$\begin{aligned} \mathbb{E}[\sum_{k=1}^K \| \sum_{t=1}^T \nabla_t^k \|_\infty] \leq & \underbrace{\mathbb{E}[\sum_{k=1}^K \sum_{t=1}^T \langle \nabla_t^k - g_t^k, x_t^k \rangle]}_{A:\text{Bias}} \\ & + \underbrace{\mathbb{E}[\sum_{k=1}^K \sum_{i=1}^d | \sum_{t=1}^T (\nabla_{t,i}^k - g_{t,i}^k) | |v_i^k|]}_{B:\text{Variance}} \\ & + \underbrace{\mathbb{E}[\sum_{k=1}^K \sum_{t=1}^T \langle g_t^k, x_t^k - v^k \rangle]}_{C:\text{Dynamic Regret}} \\ & + \frac{R}{\Delta} + \frac{2\sqrt{2} N \gamma L_1}{\Delta} \end{aligned}$$

It can be proved that g_t^k is an unbiased estimator of ∇_t^k leading to $A = 0$.

Term B is related to the variance of the estimated gradient. It follows from the properties of the subgradient of the maximum norm that the variance can be rewritten into

$$\begin{aligned} & \mathbb{E}[\sum_{k=1}^K \sum_{i=1}^d | \sum_{t=1}^T (\nabla_{t,i}^k - g_{t,i}^k) | |v_i^k|] \\ = & \sum_{k=1}^K \sum_{i=1}^d \mathbb{E}[| \sum_{t=1}^T (\nabla_{t,i}^k - g_{t,i}^k) | |v_i^k = 1] \Pr(v_i^k = 1). \end{aligned}$$

Using the property of the ℓ_1-ellipsoidal smoothing described in Lemma 1, we can prove $B \in \mathcal{O}(L_1 K (dT)^{\frac{1}{2}})$.

Finally, it follows from the property of the subdifferential of the norm that $\|\partial \|\cdot\|_\infty (g)\|_1 \leq 1$ holds for all $g \in \mathbb{R}^d$, which reduces the unconstrained stochastic optimisation to an online optimisation over $B_1(0, 1)$. Because of the ℓ_1-ellipsoidal smoothing, the per-coordinate variance of the gradient estimator is proportional to $\frac{1}{|x_{t,i}^k| + \frac{1}{d}}$, for which we apply the mirror descent method introduced in Section 3. The analysis of dynamic regret is similar to the analysis of Algorithm 1, which leads to $C \in \mathcal{O}(L_1 K (dT)^{\frac{1}{2}})$. The claimed result of Theorem 2 is obtained by choosing a large enough size of the inner loop T and a small enough smoothing parameter γ.

5. Conclusion

This paper proposes a new algorithm for bandit convex optimisation with two-point feedback over cross-polytopes. Inspired by Audibert & Bubeck, we propose an ℓ_1-ellipsoidal smoothing technique for gradient estimation and a symmetric version of the negative Tsallis-entropy for updating the decision variable, which integrates the underlying mechanism of INF seamlessly into the framework of bandit convex optimisation. The proposed algorithm guarantees a regret upper bounded by $\mathcal{O}(\sqrt{dT})$, which improves the dimensionality dependence of the best-known result (Akhavan et al., 2022) by a factor of $\sqrt{\log d}$.

Combined with the online-to-nonconvex technique proposed by Cutkosky et al., our idea can be further applied to stochastic optimisation of nonconvex and non-smooth objectives. For objective functions that are L_1-Lipschitz w.r.t 1-norm, we propose a stochastic zeroth-order algorithm, which finds a (δ, ϵ)-stationary point within $\mathcal{O}(dL_1^2\delta^{-1}\epsilon^{-3})$ function evaluations, matching the best-known result achieved recently (Kornowski & Shamir, 2023).

Despite the theoretically advantageous characteristics of the proposed algorithms, it is imperative to acknowledge the practical limitations of the idea. For gradient-free optimisation of smooth objectives with decision sets contained in cross-polytopes, the application of standard MD with mini-batches and the proposed update-generating function is straightforward. However, due to the decision variable-dependent smoothing, it is difficult to apply advanced variance reduction techniques, such as recursive gradient (Levy et al., 2021), while keeping a small dependence on dimensionality. In any case, a systematic examination of the proposed algorithm's performance through rigorous experimentation is required.

Acknowledgements

This work is supported by the German Federal Institute for Occupational Safety and Health (BAuA).

Impact Statement

This paper presents work whose goal is to advance the field of optimisation. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Agarwal, A., Dekel, O., and Xiao, L. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, June 2010.

Akhavan, A., Chzhen, E., Pontil, M., and Tsybakov, A. A gradient estimator via l1-randomization for online zero-order optimization with two point feedback. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 7685–7696. Curran Associates, Inc., 2022.

Audibert, J.-Y. and Bubeck, S. Regret bounds and minimax policies under partial monitoring. *The Journal of Machine Learning Research*, 11:2785–2836, 2010.

Chen, L., Xu, J., and Luo, L. Faster gradient-free algorithms for nonsmooth nonconvex stochastic optimization. *arXiv preprint arXiv:2301.06428*, 2023a.

Chen, L., Xu, J., and Luo, L. Faster gradient-free algorithms for nonsmooth nonconvex stochastic optimization. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5219–5233. PMLR, 23–29 Jul 2023b.

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 15–26, 2017.

Chen, P.-Y., Sharma, Y., Zhang, H., Yi, J., and Hsieh, C.-J. Ead: elastic-net attacks to deep neural networks via adversarial examples. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Choromanski, K., Rowland, M., Sindhwani, V., Turner, R., and Weller, A. Structured evolution with compact architectures for scalable policy optimization. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 970–978. PMLR, 10–15 Jul 2018.

Clarke, F. Optimization and nonsmooth analysis, siam, 1990.

Cutkosky, A., Mehta, H., and Orabona, F. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. *arXiv preprint arXiv:2302.03775*, 2023.

Duchi, J. C., Jordan, M. I., Wainwright, M. J., and Wibisono, A. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

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

Goldstein, A. Optimization of lipschitz continuous functions. *Mathematical Programming*, 13:14–22, 1977.

Kornowski, G. and Shamir, O. Oracle complexity in nonsmooth nonconvex optimization. *Journal of Machine Learning Research*, 23(314):1–44, 2022.

Kornowski, G. and Shamir, O. An algorithm with optimal dimension-dependence for zero-order nonsmooth nonconvex stochastic optimization. *arXiv preprint arXiv:2307.04504*, 2023.

Levy, K., Kavis, A., and Cevher, V. Storm+: Fully adaptive sgd with recursive momentum for nonconvex optimization. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 20571–20582. Curran Associates, Inc., 2021.

Lin, T., Zheng, Z., and Jordan, M. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26160–26175. Curran Associates, Inc., 2022.

Liu, S., Chen, J., Chen, P.-Y., and Hero, A. Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications. In *International Conference on Artificial Intelligence and Statistics*, pp. 288–297. PMLR, 2018.

Natesan Ramamurthy, K., Vinzamuri, B., Zhang, Y., and Dhurandhar, A. Model agnostic multilevel explanations. *Advances in neural information processing systems*, 33: 5968–5979, 2020.

Orabona, F. and Pál, D. Optimal non-asymptotic lower bound on the minimax regret of learning with expert advice. *arXiv preprint arXiv:1511.02176*, 2015.

Orabona, F. and Pál, D. Scale-free online learning. *Theoretical Computer Science*, 716:50–69, 2018. ISSN 0304-3975. doi: https://doi.org/10.1016/j.tcs.2017.11.021. URL `https://www.sciencedirect.com/science/article/pii/S0304397517308514`. Special Issue on ALT 2015.

Rockafellar, R. T. and Wets, R. J.-B. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Shamir, O. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *The Journal of Machine Learning Research*, 18(1):1703–1713, 2017.

Shao, W. and Albayrak, S. Adaptive zeroth-order optimisation of nonconvex composite objectives. In Nicosia, G., Ojha, V., La Malfa, E., La Malfa, G., Pardalos, P., Di Fatta, G., Giuffrida, G., and Umeton, R. (eds.), *Machine Learning, Optimization, and Data Science*, pp. 573–595, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-25599-1.

Shao, W., Sivrikaya, F., and Albayrak, S. Adaptive stochastic optimisation of nonconvex composite objectives. *arXiv preprint arXiv:2211.11710*, 2022.

Suh, H. J., Simchowitz, M., Zhang, K., and Tedrake, R. Do differentiable simulators give better policy gradients? In *International Conference on Machine Learning*, pp. 20668–20696. PMLR, 2022.

Zhang, J., Lin, H., Jegelka, S., Sra, S., and Jadbabaie, A. Complexity of finding stationary points of nonconvex nonsmooth functions. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11173–11182. PMLR, 13–18 Jul 2020a.

Zhang, J., Lin, H., Jegelka, S., Sra, S., and Jadbabaie, A. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, pp. 11173–11182. PMLR, 2020b.

A. Missing Proofs

A.1. Proof of Lemma 1

Proof of Lemma 1. The first property is obtained from the L_1-Lipschitz of f as follows

$$\begin{aligned}|f_{\gamma,\Lambda}(x) - f(x)| =&|\mathbb{E}[f(x+\gamma\Lambda u) - f(x)]|\\ \leq&\mathbb{E}[|f(x+\gamma\Lambda u) - f(x)|]\\ \leq&\mathbb{E}[L_1\|\gamma\Lambda u\|_1]\\ \leq&\gamma L_1\mathbb{E}[\|\lambda\|_2\|u\|_2]\\ \leq&\gamma\|\lambda\|_2 L_1.\end{aligned}$$

The second property is obtained by applying Lemma 1 in (Akhavan et al., 2022) to function $f(\Lambda\cdot)$ at $\Lambda^{-1}x$. Since the function $f(x+\gamma\Lambda\cdot)$ is $\gamma\|\lambda\|_2 L_1$-Lipschitz-continuous w.r.t. $\|\cdot\|_2$, we can apply Lemma 3 and Remark 1 in (Akhavan et al., 2022) and obtain the third property as follows

$$\mathbb{E}[|f(x+\gamma\Lambda u) - \mathbb{E}[f(x+\gamma\Lambda u)]|^2] \leq \frac{12\gamma^2\|\lambda\|_2^2 L_1^2}{d^2}(1+\sqrt{2})^2$$

for $d \geq 3$. □

A.2. Proof of Lemma 2

Proof of Lemma 2. The statements of differentiability at $x \neq 0$ are straightforward. For any $h > 0$, we have

$$\begin{aligned}\lim_{h\searrow 0}\frac{\psi(0+h)-\psi(0)}{h} &= \lim_{h\searrow 0}\frac{1}{h}\left(\frac{2h}{\sqrt{\beta}} - 4\sqrt{\beta+h} + 4\sqrt{\beta}\right)\\ &= \lim_{h\searrow 0}\frac{1}{h}\left(\frac{2h}{\sqrt{\beta}} - \frac{4h}{\sqrt{\beta}+\sqrt{\beta+h}}\right)\\ &=0\end{aligned}$$

Analogously, we also have $\lim_{h\nearrow 0}\frac{\psi(0+h)-\psi(0)}{h} = 0$, which implies $\psi'(0) = 0$. Using the same argument, it can be shown that ψ is twice differentiable at 0 with $\psi''(0) = \beta^{-\frac{3}{2}}$. Since $\psi''(x) > 0$ for all $x \in \mathbb{R}$, we obtain the strict convexity.

Let $\theta \in (-\frac{2}{\sqrt{\beta}}, \frac{2}{\sqrt{\beta}})$ be in the range of ψ' with $\theta = \psi'(x)$. Since $\operatorname{sgn}(\psi'(x)) = \operatorname{sgn}(x)$ holds for all x, the sign of θ is same as the sign of x. Thus, we take the inverse of ψ' according to the sign of θ.

To prove the last property, we first note that there is a $c \in (0,1)$ such that

$$\begin{aligned}\mathcal{B}_\psi(y,x) =&\frac{1}{2}(|cx+(1-c)y|+\beta)^{-\frac{3}{2}}(x-y)^2\\ \geq&\frac{1}{2}(\max\{|x|,|y|\}+\beta)^{-\frac{3}{2}}(x-y)^2.\end{aligned}$$

W.l.o.g. we assume $|\psi'(x)| \leq |\psi'(y)|$. Define $g = \psi'(x) - \psi'(y)$. By the assumption on g, $\psi'(x) + \operatorname{sgn}(\psi'(x))|g|$ is in the range of ψ'. Therefore, there is a z satisfying $\psi'(z) = \psi'(x) + \operatorname{sgn}(\psi'(x))|g|$. Since ψ' is increasing,

$$|\psi'(z)| = |\psi'(x) + \operatorname{sgn}(\psi'(x))|g|| = |\psi'(x)| + |g| \geq |\psi'(x) - g| = |\psi'(y)|$$

implies $|z| \geq |y|$ and

$$\begin{aligned}\mathcal{B}_\psi(y,x)\ \geq&\frac{1}{2}(\max\{|x|,|y|\}+\beta)^{-\frac{3}{2}}(x-y)^2\\ \geq&\frac{1}{2}(|z|+\beta)^{-\frac{3}{2}}(x-y)^2.\end{aligned}$$

Since z and x have the same sign, we have for $\beta \leq 1$

$$|z|+\beta = \frac{4}{(\frac{2}{\sqrt{\beta+|x|}} - |g|)^2} = \frac{4(|x|+\beta)}{(2-|g|\sqrt{|x|+\beta})^2}.$$

For $0 \leq |g|\sqrt{|x|+\beta} \leq 1$, we have $2 - |g|\sqrt{|x|+\beta} \geq 1$ and

$$|z| + \beta \leq 4(|x| + \beta).$$

Thus, we obtain the last property

$$\begin{aligned}\mathcal{B}_\psi(y,x) \geq& \frac{1}{2}(|z|+\beta)^{-\frac{3}{2}}(x-y)^2\\ \geq& \frac{1}{16}(|x|+\beta)^{-\frac{3}{2}}(x-y)^2.\end{aligned}$$

□

A.3. Proof of Lemma 3

Proof of Lemma 3. First, we obtain the following inequality from the optimality condition of the update rule.

$$\begin{aligned}\langle g_t, x_{t+1} - x\rangle \leq& \alpha_t\langle \nabla\phi(x_{t+1}) - \nabla\phi(x_t), x - x_{t+1}\rangle\\ =& \alpha_t\mathcal{B}_\phi(x,x_t) - \alpha_t\mathcal{B}_\phi(x,x_{t+1}) - \alpha_t\mathcal{B}_\phi(x_{t+1},x_t).\end{aligned} \tag{8}$$

Following the standard procedure of analysing MD, we have

$$\begin{aligned}\langle g_t, x_t - x\rangle =& \langle g_t, x_t - x_{t+1}\rangle + \langle g_t, x_{t+1} - x\rangle\\ \leq& \langle g_t, x_t - x_{t+1}\rangle + \alpha_t\mathcal{B}_\phi(x,x_t) - \alpha_t\mathcal{B}_\phi(x,x_{t+1}) - \alpha_t\mathcal{B}_\phi(x_{t+1},x_t)\\ \leq& \alpha_t\mathcal{B}_\phi(x,x_t) - \alpha_t\mathcal{B}_\phi(x,x_{t+1}) + \frac{4}{\alpha_t}\sum_{i=1}^d(|x_{t,i}|+\beta)^{\frac{3}{2}}|g_{t,i}|^2\end{aligned} \tag{9}$$

where the first inequality uses (8) and the second inequality follows from the assumption $|g_{t,i}|\sqrt{|x_{t,i}|+\beta} \leq \alpha_t$ and Lemma 2. W.l.o.g. we assume $\alpha_0 = 0$. Adding up from 1 to T, we obtain

$$\begin{aligned}\sum_{t=1}^T\langle g_t, x_t - x\rangle \leq& \sum_{t=1}^T\alpha_t(\mathcal{B}_\phi(x,x_t) - \mathcal{B}_\phi(x,x_{t+1})) + \sum_{t=1}^T\frac{4}{\alpha_t}\sum_{i=1}^d(|x_{t,i}|+\beta)^{\frac{3}{2}}|g_{t,i}|^2\\ \leq& \alpha_0\mathcal{B}_\phi(x,x_1) + \sum_{t=1}^T(\alpha_t - \alpha_{t-1})\mathcal{B}_\phi(x,x_t) + \sum_{t=1}^T\frac{4}{\alpha_t}\sum_{i=1}^d(|x_{t,i}|+\beta)^{\frac{3}{2}}|g_{t,i}|^2\\ =& \alpha_T\max_{x,y\in\mathcal{K}}\mathcal{B}_\phi(x,y) + \sum_{t=1}^T\frac{4}{\alpha_t}\sum_{i=1}^d(|x_{t,i}|+\beta)^{\frac{3}{2}}|g_{t,i}|^2.\end{aligned} \tag{10}$$

The Bregman divergence can be upper bounded by

$$\begin{aligned}\mathcal{B}_\phi(x,y) \leq& \langle\nabla\phi(x) - \nabla\phi(y), x - y\rangle\\ \leq& \|\nabla\phi(x) - \nabla\phi(y)\|_\infty\|x-y\|_1,\\ \leq& 8D\beta^{-\frac{1}{2}},\end{aligned} \tag{11}$$

where the first inequality follows from the convexity of ϕ, the second line uses the Hölder's inequality and the last line uses the boundness of the decision set. Combining (10) and (11), we obtain the desired result. □

A.4. Proof of Proposition 1

Proof of Proposition 1. The proof follows the idea in (Lin et al., 2022). Define $\tilde{f}(x) : \mathbb{R}^d \to \mathbb{R}, x \mapsto f(\Lambda x)$. Since f is Lipschitz continuous, $\tilde{f}$ is also Lipschitz continuous. It follows from Lemma 1 in (Akhavan et al., 2022) that $\tilde{f}_\gamma := \mathbb{E}_u[\tilde{f}(\cdot + \gamma u)]$ is differentiable everywhere. Thus, for all $x \in \mathbb{R}^d$, we have

$$\begin{aligned}&\lim_{\|\Lambda^{-1}v\|_2\to 0}\frac{f_{\gamma,\Lambda}(x+v) - f_{\gamma,\Lambda}(x) - \langle\Lambda^{-1}\nabla\tilde{f}_\gamma(\Lambda^{-1}x), v\rangle}{\|\Lambda^{-1}v\|_2}\\ =&\lim_{\|\Lambda^{-1}v\|_2\to 0}\frac{\tilde{f}_\gamma(\Lambda^{-1}x + \Lambda^{-1}v) - \tilde{f}_\gamma(\Lambda^{-1}x) - \langle\nabla\tilde{f}_\gamma(\Lambda^{-1}x), \Lambda^{-1}v\rangle}{\|\Lambda^{-1}v\|_2}\\ =&0,\end{aligned} \tag{12}$$

which implies that $f_{\gamma,\Lambda}$ is differentiable with $\nabla f_{\gamma,\Lambda}(x) = \Lambda^{-1}\nabla \tilde{f}_\gamma(\Lambda^{-1}x)$. It follows from the Lipschitz continuity of f that f is almost everywhere differentiable. Let $U \subseteq B_1(0,\gamma)$ with $\mathrm{vol}(B_1(0,\gamma)) = \mathrm{vol}(U)$ and f is everywhere differentiable in U. For a differentiable vector $x + \Lambda u$ and any h, the L_1-Lipschitz of f ensures

$$|\frac{f(x+h+\Lambda u) - f(x+\Lambda u) - \langle \nabla f(x+\Lambda u), h\rangle}{\|h\|_1}| \le 2L_1.$$

It follows from the dominated convergence theorem that

$$\begin{aligned}
&\lim_{\|h\|_1\to 0} \frac{f_{\gamma,\Lambda}(x+h) - f_{\gamma,\Lambda}(x) - \langle \frac{1}{\mathrm{vol}(U)}\int_{u\in U}\nabla f(x+\Lambda u)du, h\rangle}{\|h\|_1}\\
=&\lim_{\|h\|_1\to 0} \frac{f_{\gamma,\Lambda}(x+h) - f_{\gamma,\Lambda}(x) - \frac{1}{\mathrm{vol}(B_1(0,\gamma))}\int_{u\in U}\langle\nabla f(x+\Lambda u), h\rangle du}{\|h\|_1}\\
=&\frac{1}{\mathrm{vol}(B_1(0,\gamma))}\int_{u\in U}(\lim_{\|h\|_1\to 0}\frac{f(x+h+\Lambda u) - f(x+\Lambda u) - \langle\nabla f(x+\Lambda u), h\rangle}{\|h\|_1})du\\
=&0,
\end{aligned}$$

holds for all x, i.e. $\nabla f_{\gamma,\Lambda}(x) = \frac{1}{\mathrm{vol}(B_1(0,\gamma))}\int_{u\in U}(\nabla f(x+\Lambda u))du$.

Next, we define the set $U_\Lambda(x) = \{x + \Lambda u | u \in U\}$. Then, the subgradient can be rewritten into

$$\nabla f_{\gamma,\Lambda}(x) = \frac{1}{\mathrm{vol}(B_1(0,\gamma))}\int_{y\in U_\Lambda(x)}(\nabla f(y))dy$$

For $\Lambda = \mathrm{diag}(\lambda)$ and any $y \in U_\Lambda(x)$, there is an u in U such that

$$\|y-x\|_1 \le \|\Lambda u\|_1 \le \|\lambda\|_2\|u\|_2 \le \|\lambda\|_2\|u\|_1 \le \gamma\|\lambda\|_2,$$

i.e. $U_\Lambda(x) \subseteq B_1(x, \gamma\|\lambda\|_2)$. Assume the existence of x_0 with $\nabla f_{\gamma,\Lambda}(x_0) \notin \partial_{\gamma\|\lambda\|_2} f(x_0)$. Then there is some $w \in \mathbb{R}^d$ and $c \in \mathbb{R}$ such that

$$\langle w, g\rangle < c \le \langle w, \nabla f_{\gamma,\Lambda}(x_0)\rangle,$$

for all $g \in \partial_{\gamma\|\lambda\|_2} f(x_0)$ (Rockafellar & Wets, 2009), which leads to a contradiction to $U_\Lambda(x) \subseteq B_1(x, \gamma\|\lambda\|_2)$. Thus, we conclude $\nabla f_{\gamma,\Lambda}(x) \in \partial_{\gamma\|\lambda\|_2} f(x)$ for all $x \in \mathbb{R}^d$.

Finally, let $g \in \partial_\delta f_{\gamma,\Lambda}(x)$ be in the δ-subdifferential of $f_{\gamma,\Lambda}$ at some point x. Then, by definition, there are $y_1, \dots, y_n \in B_1(x,\delta)$ and $a \in [0,1]^n$ such that $g = \sum_{i=1}^n a_i \nabla f_{\gamma,\Lambda}(y_i)$. Since

$$\nabla f_{\gamma,\Lambda}(y_i) \in \partial_{\gamma\|\lambda\|_2} f(y_i) = \mathrm{conv}\{\cup_{z\in B_1(y_i,\gamma\|\lambda\|_2)}\partial f(z)\} \subseteq \mathrm{conv}\{\cup_{z\in B_1(x,\delta+\gamma\|\lambda\|_2)}\partial f(z)\} = \partial_{\delta+\gamma\|\lambda\|_2} f(x)$$

holds for all $i = 1\dots, n$, we obtain $g \in \partial_{\delta+\gamma\|\lambda\|_2} f(x)$. □

A.5. Proof of Theorem 2

Proof of Lemma 4. For simplicity, we set $N = KT$ and we use $v_t^k = v_{(k-1)T+t}$ for any sequence $v_1, \dots, v_{KT}$. First of all, we have

$$\begin{aligned}
&F(y_{n+1}) - F(y_n)\\
=&F_{\gamma,\Lambda_n}(y_{n+1}) - F_{\gamma,\Lambda_n}(y_n) + F(y_{n+1}) - F_{\gamma,\Lambda_n}(y_{n+1}) - F(y_n) + F_{\gamma,\Lambda_n}(y_n)\\
\le&F_{\gamma,\Lambda_n}(y_{n+1}) - F_{\gamma,\Lambda_n}(y_n) + 2\gamma\|\lambda_n\|_2 L_1\\
=&\int_0^1 \langle\nabla F_{\gamma,\Lambda_n}(y_n + \eta\Delta x_n), \Delta x_n\rangle d\eta + 2\gamma\|\lambda_n\|_2 L_1\\
=&\langle\mathbb{E}[\nabla F_{\gamma,\Lambda_n}(y_t + \eta\Delta x_n)], \Delta x_n\rangle + 2\gamma\|\lambda_n\|_2 L_1\\
=&\langle\nabla_n, \Delta x_n\rangle + 2\gamma\|\lambda_n\|_2 L_1\\
=&\Delta\langle\nabla_n - g_n, x_n\rangle + \Delta\langle g_n, x_n - w_n\rangle + \Delta\langle g_n, w_n\rangle + 2\gamma\|\lambda_n\|_2 L_1,
\end{aligned}$$

where the first inequality follows from Lemma 2, the third line uses the differentability of $f_{\gamma,\Lambda}$, we define $\triangledown_n = \mathbb{E}[\triangledown f_{\gamma,\Lambda_n}(z_n + \eta x_n)]$ and $\{w_n\}$ is any sequence. Adding up from 1 to N, taking expectation and rearranging, we obtain

$$\begin{aligned} -R \le & F(y_{N+1}) - F(y_1)\\ =& \sum_{n=1}^{N}(F(y_{t+1}) - F(y_t))\\ \le & \Delta\sum_{n=1}^{N}\langle \triangledown_n - g_n, x_n\rangle + \Delta\sum_{n=1}^{N}\langle g_n, x_n - w_n\rangle + \Delta\sum_{n=1}^{N}\langle g_n, w_n\rangle + 2\gamma L_1\sum_{n=1}^{N}\|\lambda_n\|_2 \end{aligned} \tag{13}$$

Combining (13) and (14) and rearranging, we have

$$-\sum_{n=1}^{N}\langle g_n, w_n\rangle \le \sum_{n=1}^{N}\langle \triangledown_n - g_n, x_n\rangle + \frac{R}{\Delta} + \sum_{n=1}^{N}\langle g_n, x_n - w_n\rangle + \frac{2\gamma L_1\sum_{n=1}^{N}\|\lambda_n\|_2}{\Delta}.$$

Setting $w_1^k =, \dots, = w_T^k = v^k$ and using the fact $\|\lambda_n\|_2 \le \sqrt{2}$, we obtain the claimed result. □

Proof of Theorem 2. For simplicity, we set $N = KT$ and we use $v_t^k = v_{(k-1)T+t}$ for any sequence $v_1, \dots, v_{KT}$. Lemma 4 can be directly applied with $-w_1^k =, \dots, = -w_T^k = -v^k \in \partial\|\cdot\|_\infty(\sum_{t=1}^{T}\triangledown_t^k)$ and yields

$$-\sum_{k=1}^{K}\sum_{t=1}^{T}\langle g_t^k, v^k\rangle \le \sum_{k=1}^{K}\sum_{t=1}^{T}\langle \triangledown_t^k - g_t^k, x_t^k\rangle + \frac{R}{\Delta} + \sum_{k=1}^{K}\sum_{t=1}^{T}\langle g_t^k, x_t^k - v^k\rangle + \frac{2\sqrt{2}\gamma L_1 N}{\Delta}.$$

With the choice of v^k and the property of the subgradient of norms, we obtain

$$\begin{aligned} -\sum_{k=1}^{K}\sum_{t=1}^{T}\langle g_t^k, v^k\rangle =& \sum_{k=1}^{K}\langle\sum_{t=1}^{T}(\triangledown_t^k - g_t^k), v_k\rangle - \sum_{k=1}^{K}\langle\sum_{t=1}^{T}\triangledown_t^k, v_k\rangle\\ =& \sum_{k=1}^{K}\langle\sum_{t=1}^{T}(\triangledown_t^k - g_t^k), v_k\rangle + \sum_{k=1}^{K}\|\sum_{t=1}^{T}\triangledown_t^k\|_\infty\\ \ge & -\sum_{k=1}^{K}\sum_{i=1}^{d}|\sum_{t=1}^{T}(\triangledown_{t,i}^k - g_{t,i}^k)||v_i^k| + \sum_{k=1}^{K}\|\sum_{t=1}^{T}\triangledown_t^k\|_\infty. \end{aligned}$$

Combining the inequality above, rearranging and taking expectation, we obtain

$$\begin{aligned} \mathbb{E}[\sum_{k=1}^{K}\|\sum_{t=1}^{T}\triangledown_t^k\|_\infty] \le & \frac{R}{\Delta} + \frac{2\sqrt{2}N\gamma L_1}{\Delta} + \underbrace{\mathbb{E}[\sum_{k=1}^{K}\sum_{t=1}^{T}\langle \triangledown_t^k - g_t^k, x_t^k\rangle]}_{A\text{:Bias}}\\ & + \underbrace{\mathbb{E}[\sum_{k=1}^{K}\sum_{i=1}^{d}|\sum_{t=1}^{T}(\triangledown_{t,i}^k - g_{t,i}^k)||v_i^k|]}_{B\text{:Variance}} + \underbrace{\mathbb{E}[\sum_{k=1}^{K}\sum_{t=1}^{T}\langle g_t^k, x_t^k - v^k\rangle]}_{C\text{:Dynamic Regret}} \end{aligned}$$

Bounding A We first show that g_n is an unbiased estimator of $\triangledown_n$. Using the total law of expectation and the definition of g_n, we have

$$\begin{aligned} & \mathbb{E}[\langle \triangledown_n - g_n, x_n\rangle]\\ =& \mathbb{E}[\mathbb{E}_{\eta_n,u_n,\xi_n}[\mathbb{E}_{u_n,\xi_n}[\mathbb{E}_{\xi_n}[\langle \triangledown_n - g_n, x_n\rangle|u_n]|\eta_n]|x_n, y_n]]\\ =& \mathbb{E}[\mathbb{E}_{\eta_n,u_n,\xi_n}[\mathbb{E}_{u_n,\xi_n}[\langle \triangledown_n - \frac{d}{2\gamma}(F(y_n + \eta_n x_n + \gamma\Lambda_n u_n) - F(y_n + \eta_n x_n - \gamma\Lambda_n u_n))\Lambda_n^{-1}\operatorname{sgn}(u_n), x_n\rangle|\eta_n]|x_n, y_n]]\\ =& \mathbb{E}[\mathbb{E}_{\eta_n,u_n,\xi_n}[\langle \triangledown_n - \triangledown F_{\gamma,\Lambda_n}(y_n + \eta_n x_n), x_n\rangle|x_n, y_n]]\\ =& 0. \end{aligned} \tag{14}$$

Bounding B To upper bound the variance, we consider a special choice of v_k. Define $I = \arg\max_{i\in[d]} |\sum_{t=1}^T \nabla_{t,i}^k|$. We choose

$$v_i^k = \begin{cases} \frac{\operatorname{sgn}(\sum_{t=1}^T \nabla_{t,k}^k)}{|I|} & i \in I \\ 0 & i \notin I, \end{cases}$$

v^k is clearly a subdifferential of the maximum norm. Furthermore, if $\sum_{t=1}^T \nabla_t^k$ is a random vector, v^k is also a random vector. Thus, we obtain

$$\mathbb{E}[\sum_{k=1}^K \sum_{i=1}^d |\sum_{t=1}^T (\nabla_{t,i}^k - g_{t,i}^k)||v_i^k|] = \sum_{k=1}^K \sum_{i=1}^d \mathbb{E}[|\sum_{t=1}^T (\nabla_{t,i}^k - g_{t,i}^k)||v_i^k = 1] \Pr(v_i^k = 1).$$

Since g_t^k is an unbiased estimator of ∇_t^k, the conditional variance can be bounded as follows.

$$\begin{aligned}
\mathbb{E}[|\sum_{t=1}^T (\nabla_{t,i}^k - g_{t,i}^k)|^2|v_i^k = 1] =& \sum_{t=1}^T \mathbb{E}[(\nabla_{t,i}^k - g_{t,i}^k)^2|v_i^k = 1] \\
\leq& \sum_{t=1}^T \mathbb{E}[(g_{t,i}^k)^2|v_i^k = 1] \\
=& \sum_{t=1}^T \mathbb{E}[\mathbb{E}[\mathbb{E}[(g_{t,i}^k)^2|z_t]|\xi_t^k]|v_i^k = 1] \\
\leq& \frac{d^3}{4\gamma^2} \sum_{t=1}^T \mathbb{E}[\mathbb{E}[\mathbb{E}[|f(z_t^k + \gamma\Lambda_t^k u_t^k, \xi_t^k) - f(z_t^k - \gamma\Lambda_t^k u_t^k, \xi_t^k)|^2|z_t]|\xi_t^k]|v_i^k = 1] \\
\leq& 2cd \sum_{t=1}^T \mathbb{E}[L_1(\xi_t^k)^2|v_i^k = 1] \\
\leq& 2cdTL_1^2,
\end{aligned}$$

where the first inequality follows from the property of variance, the second inequality uses the definition of g_n and choice of β, the third inequality follows from Lemma 1 and the last inequality uses Assumption 1. Applying Jensen's inequality and total law of expectation, we obtain

$$\begin{aligned}
\mathbb{E}[\sum_{k=1}^K \langle \sum_{t=1}^T (\nabla_t^k - g_t^k), v_k \rangle] \leq& \sum_{k=1}^K \sum_{i=1}^d \mathbb{E}[|\sum_{t=1}^T (\nabla_{t,i}^k - g_{t,i}^k)||v_i^k = 1] \Pr(v_i^k = 1) \\
=& \sum_{k=1}^K \sum_{i=1}^d \mathbb{E}[(|\sum_{t=1}^T (\nabla_{t,i}^k - g_{t,i}^k)|^2)^{\frac{1}{2}}|v_i^k = 1] \Pr(v_i^k = 1) \\
\leq& \sum_{k=1}^K \sum_{i=1}^d (\mathbb{E}[|\sum_{t=1}^T (\nabla_{t,i}^k - g_{t,i}^k)|^2|v_i^k = 1])^{\frac{1}{2}} \Pr(v_i^k = 1) \\
\leq& \sum_{k=1}^K (2cdT)^{\frac{1}{2}} L_1 \sum_{i=1}^d \Pr(v_i^k = 1) \\
=& (2cdT)^{\frac{1}{2}} K L_1.
\end{aligned}$$

Bounding C Finally, the dynamic regret can be upper bounded by applying Lemma 3 to each $k \in [K]$. Setting

$$\alpha_t^k = (\frac{d^2}{4\gamma^2} \sum_{s=1}^t |f(z_s^k + \gamma\Lambda_s^k u_s^k) - f(z_s^k - \gamma\Lambda_s^k u_s^k)|^2)^{\frac{1}{2}}$$

ensures $|g_{t,i}^k|\sqrt{|x_{t,i}^k| + \beta} \leq \alpha_t^k$. From Lemma 3 it follows that

$$\begin{aligned}
\sum_{t=1}^{T}\langle g_t^k, x_t^k - v^k\rangle \leq & 8\alpha_T^k d^{\frac{1}{2}} + \sum_{t=1}^{T}\frac{4}{\alpha_t^k}\sum_{i=1}^{d}(|x_{t,i}^k| + \beta)^{\frac{3}{2}}|g_{t,i}^k|^2 \\
= & 8\alpha_T^k d^{\frac{1}{2}} + \sum_{t=1}^{T}\frac{4}{\alpha_t^k}\frac{d^2}{4\gamma^2}|f(z_t^k + \gamma\Lambda_t^k u_t^k, \xi_t^k) - f(z_t^k - \gamma\Lambda_t^k u_t^k, \xi_t^k)|^2\sum_{i=1}^{d}(|x_{t,i}^k| + \beta)^{\frac{1}{2}} \\
\leq & 8\alpha_T^k d^{\frac{1}{2}} + d^{\frac{1}{2}}\sum_{t=1}^{T}\frac{4}{\alpha_t^k}\frac{d^2}{4\gamma^2}|f(z_t^k + \gamma\Lambda_t^k u_t^k, \xi_t^k) - f(z_t^k - \gamma\Lambda_t^k u_t^k, \xi_t^k)|^2\sqrt{\sum_{i=1}^{d}(|x_{t,i}^k| + \beta)} \\
\leq & 8\alpha_T^k d^{\frac{1}{2}} + 4\sqrt{2}d^{\frac{1}{2}}\sum_{t=1}^{T}\frac{1}{\alpha_t^k}\frac{d^2}{4\gamma^2}|f(z_t^k + \gamma\Lambda_t^k u_t^k, \xi_t^k) - f(z_t^k - \gamma\Lambda_t^k u_t^k, \xi_t^k)|^2 \\
\leq & (8d^{\frac{1}{2}} + 8\sqrt{2}d^{\frac{1}{2}})(\sum_{t=1}^{T}\frac{d^2}{4\gamma^2}|f(z_t^k + \gamma\Lambda_t^k u_t^k, \xi_t^k) - f(z_t^k - \gamma\Lambda_t^k u_t^k, \xi_t^k)|^2)^{\frac{1}{2}}.
\end{aligned}$$

Taking expectations, we have

$$\begin{aligned}
\mathbb{E}[\sum_{t=1}^{T}\langle g_t^k, x_t^k - v^k\rangle] \leq & (8d^{\frac{1}{2}} + 8\sqrt{2}d^{\frac{1}{2}})\mathbb{E}[(\sum_{t=1}^{T}\frac{d^2}{4\gamma^2}|f(z_t^k + \gamma\Lambda_t^k u_t^k, \xi_t^k) - f(z_t^k - \gamma\Lambda_t^k u_t^k, \xi_t^k)|^2)^{\frac{1}{2}}] \\
\leq & (8d^{\frac{1}{2}} + 8\sqrt{2}d^{\frac{1}{2}})(\sum_{t=1}^{T}\mathbb{E}[\frac{d^2}{4\gamma^2}|f(z_t^k + \gamma\Lambda_t^k u_t^k, \xi_t^k) - f(z_t^k - \gamma\Lambda_t^k u_t^k, \xi_t^k)|^2])^{\frac{1}{2}}
\end{aligned}$$

Applying Lemma 1, we obtain

$$\begin{aligned}
& \mathbb{E}[(\frac{d^2}{4\gamma^2}|f(x_t^k + \gamma\Lambda_t^k u_t^k, \xi_t^k) - f(x_t^k - \gamma\Lambda_t^k u_t^k, \xi_t^k)|^2] \\
= & \mathbb{E}[\mathbb{E}[\mathbb{E}[\frac{d^2}{4\gamma^2}|f(x_t^k + \gamma\Lambda_t^k u_t^k, \xi_t^k) - f(x_t^k - \gamma\Lambda_t^k u_t^k, \xi_t^k)|^2|\xi_t^k]|x_t^k]] \\
\leq & \frac{d^2}{\gamma^2}\mathbb{E}[\mathbb{E}[\mathrm{Var}(f(x_t^k + \gamma\Lambda_t^k u_t^k, \xi_t^k)|\xi_t^k)|x_t^k]] \\
\leq & 2c\mathbb{E}[\mathbb{E}[L_1(\xi_t^k)^2|x_t^k]] \\
\leq & 2cL_1^2,
\end{aligned}$$

where the second inequality follows from Lemma 1 and the last inequality uses Assumption 1. Thus, the dynamic regret is upper bounded by

$$\mathbb{E}[\sum_{t=1}^{T}\langle g_t^k, x_t^k - v^k\rangle] \leq (8\sqrt{2} + 16)c^{\frac{1}{2}}d^{\frac{1}{2}}L_1T^{\frac{1}{2}}$$

Combining the upper bounds of A, B and C, we have

$$\mathbb{E}[\sum_{k=1}^{K}\|\sum_{t=1}^{T}\triangledown_t^k\|_\infty] \leq \frac{R}{\Delta} + \frac{2\sqrt{2}N\gamma L_1}{\Delta} + (2cdT)^{\frac{1}{2}}KL_1 + (8\sqrt{2} + 16)c^{\frac{1}{2}}d^{\frac{1}{2}}L_1KT^{\frac{1}{2}}.$$

Finally, $\triangledown_t^k \in \partial_{T\Delta}F_{\gamma,\Lambda_t^k}(\bar{z}^k) \subseteq \partial_{T\Delta+2\gamma}F(\bar{z}^k)$ follows definition of δ-subdifferential and Proposition 1. Thus, we have $\frac{1}{T}\sum_{t=1}^{T}\triangledown_t^k \in \partial_{T\Delta+2\gamma}F(\bar{z}^k)$ and

$$\begin{aligned}
\mathbb{E}[\|\triangledown F(\bar{z})\|_{T\Delta+2\gamma}] \leq & \mathbb{E}[\frac{1}{K}\sum_{k=1}^{K}\|\frac{1}{T}\sum_{t=1}^{T}\triangledown_t^k\|_\infty] \\
\leq & \frac{R}{N\Delta} + \frac{2\sqrt{2}\gamma L_1}{\Delta} + \frac{(2cd)^{\frac{1}{2}}L_1}{T^{\frac{1}{2}}} + \frac{(8\sqrt{2} + 16)c^{\frac{1}{2}}d^{\frac{1}{2}}L_1}{T^{\frac{1}{2}}}.
\end{aligned}$$

Setting $T = \min\{\frac{N}{2}, (N\delta d^{\frac{1}{2}} L_1 R^{-1})^{\frac{2}{3}}\}$, $\Delta = \frac{\delta}{2T}$,$\gamma = \min\{\frac{\delta}{4}, T^{-1} N^{-\frac{1}{3}} \delta^{\frac{2}{3}} (RdL_1^2)^{\frac{1}{3}}\}$, we obtain

$$\mathbb{E}[\|\triangledown F(\bar{z})\|_\delta] \leq \max\{N^{-\frac{1}{2}} \bar{c}_3 d^{\frac{1}{2}} L_1, \tilde{c}_3 (N\delta)^{-\frac{1}{3}} (RdL_1^2)^{\frac{1}{3}}\},$$

where $\bar{c}_3$ and $\tilde{c}_3$ care constants independent of d, L_1 and R. Thus, given a fixed ϵ and δ, the algorithm requires $\mathcal{O}(\frac{RdL_1^2}{\epsilon^3 \delta})$ steps to ensure $\mathbb{E}[\|\triangledown F(\bar{z})\|_\delta] \leq \epsilon$. □