Black-Box Uniform Stability for Non-Euclidean Empirical Risk Minimization

Simon Vary
University of Oxford

David Martínez-Rubio
Carlos III University of Madrid

Patrick Rebeschini
University of Oxford

Abstract

We study first-order algorithms that are uniformly stable for empirical risk minimization (ERM) problems that are convex and smooth with respect to p-norms, $p \geq 1$. We propose a black-box reduction method that, by employing properties of uniformly convex regularizers, turns an optimization algorithm for Hölder smooth convex losses into a uniformly stable learning algorithm with optimal statistical risk bounds on the excess risk, up to a constant factor depending on p. Achieving a black-box reduction for uniform stability was posed as an open question by Attia and Koren (2022), which had solved the Euclidean case $p = 2$. We explore applications that leverage non-Euclidean geometry in addressing binary classification problems.

1 INTRODUCTION

We study how to obtain an *optimally* stable algorithm for a general learning problem from an optimization algorithm in a black-box manner. Given a distribution P within a family $\mathcal{P}$, our task is to minimize the *population risk* $f(x) = \mathbb{E}_{z\sim P}[\ell(x; z)]$ where $\ell(\cdot, z) : \mathcal{X} \to \mathbb{R}$ is a convex smooth loss function for every z. In this work, we focus on convex smooth losses with respect to p-norms. The statistical question is how to bound the *excess risk* of an estimator $\hat{x}$, which is measured as the difference between the population risk of $\hat{x}$ and the minimal population risk over the parameter domain $\mathcal{X}$:

$$\delta f(\hat{x}) = f(\hat{x}) - f(\tilde{x}),$$

where $\tilde{x} \in \arg\min_{x\in\mathcal{X}} f(x)$.

In the classical situation where the data distribution is unknown and we have access only to a finite training dataset $S = \{z_1, \dots, z_n\} \subset \mathcal{Z}$ of n-samples drawn i.i.d. from P, we can use the *empirical risk* $f_S(x) = \frac{1}{n}\sum_{i=1}^n \ell(x; z_i)$ as a sample-average proxy of the population risk. The expected excess risk of an estimator is typically bounded by balancing the trade-off between error terms originating from the generalization component and those arising from offline optimization of the empirical risk:

$$\mathbb{E}_S\left[\delta f(\hat{x})\right] = \underbrace{\mathbb{E}_S\left[f(\hat{x}) - f_S(\hat{x})\right]}_{\text{generalization}} + \underbrace{\mathbb{E}_S\left[f_S(\hat{x}) - f_S(\tilde{x})\right]}_{\text{optimization}}. \tag{1}$$

Here the optimization error can be further upper bounded using the empirical risk minimizer (ERM) $x^* \in \arg\min_{\mathcal{X}} f_S(x)$ and noting that $\mathbb{E}_S\left[f_S(\hat{x}) - f_S(\tilde{x})\right] \leq \mathbb{E}_S\left[f_S(\hat{x}) - f_S(x^*)\right]$.

In this paper we are interested in providing optimal risk bounds which involve bounding both the generalization and the optimization error. The notion of *algorithmic stability* (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2009) has been successfully used to control the generalization error. This notion pertains to the ability of the algorithm to be robust to small perturbations in the training dataset.

Using the *uniform stability* framework, which is a particular type of algorithmic stability, Attia and Koren (2022) showed that it is possible to perform a black-box conversion from a given optimization algorithm for a convex, smooth objective to a uniformly stable algorithm with the same optimization convergence rate in the Euclidean case ($p = 2$), up to a log factor. However, Attia and Koren (2022) noted that it is currently unknown whether similar black-box results are achievable when the function regularity is measured with general non-Euclidean geometries $\|\cdot\|_p$ for $p \geq 1$, posing the following open problem:

> *Given an optimization algorithm for a convex, and smooth objective in ℓ_p-geometry for $p \geq 1$, is it possible to perform a black-box conversion to an algorithm with the same convergence rate that is also uniformly stable?*

1.1 Contributions

In this paper, motivated by the open problem of Attia and Koren (2022), we develop techniques for uniformly stable

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

empirical risk minimization with convex smooth objectives with respect to non-Euclidean norms $\|\cdot\|_p$. By studying the uniform stability of estimators that are approximate minimizers of the empirical risk minimization after adding uniformly convex regularization, see Lemma 5, we design uniformly stable algorithms for problems with regularity measured in general $\|\cdot\|_p$ normed spaces.

Black-box reduction in non-Euclidean geometry. We design an algorithm, *Uniform Stable Optimization with ℓ_p Regularity* (USOLP), see Algorithm 1. The algorithm performs a black-box reduction from a given optimization algorithm for convex Hölder smooth functions that is assumed to remain within a bounded domain to a learning algorithm with the same convergence rate that is *optimally uniformly stable* (Theorem 8). Our black-box reduction to a stable algorithm is based on: (A) stability properties of uniformly convex regularization (Lemma 5), and (B) observing that restarting an optimization algorithm for a convex Hölder smooth functions captures the *uniformly convex structure* of the objective and achieves corresponding convergence rates, which can be faster in some situations, e.g., see (Renegar and Grimmer, 2022). Thus we provide a positive answer to the open problem posed by Attia and Koren (2022) with the caveat that our USOLP reduction algorithm for ℓ_p geometry works for optimization methods for $(L, \min\{2,p\})$-Hölder smooth objectives and requires knowledge of an upper bound on the distance of the initial point to the minimizer, the Lipschitz constant, and the guarantee that the given optimization method remains in the corresponding domain. The reduction is also optimal, since it exhibits the best convergence versus stability trade-off: an improvement to either of the rates would contradict the statistical lower bounds that were derived by Levy and Duchi (2019) for $p \in [1,2]$, and that we establish for $p \geq 2$ in Theorem 7.

Optimal expected excess risk bounds. Theorem 6 shows that an (approximate) minimizer of the empirical risk with a specific level of uniformly convex regularization achieves an optimal expected excess risk, up to a constant factor. Consequently, Corollary 10 and Corollary 11 show that our black-box reductions converge to a point x_T that achieves an excess risk

$$\mathbb{E}_S\left[\delta f(x_T)\right] = \begin{cases} \mathcal{O}_p\left(LR^2 \frac{d^{1/2-1/\hat{p}}}{n^{1/2}}\right) & \text{when } d \leq n \\ \mathcal{O}_p\left(LR^2 \frac{1}{n^{1/\hat{p}}}\right) & \text{when } d > n \end{cases},$$

that is optimal up to a constant factor in p for L-smooth losses over the domain $\mathcal{B}_{\|\cdot\|_p}(x_0, R)$ where x_0 is the starting point of our algorithm and $\hat{p} = \max\{2,p\}$. These rates are identical up to a constant to the excess risk bound obtained by employing the *non*-black-box algorithm proposed by (Diakonikolas and Guzmán, 2024), see Corollary 11 and Corollary 13 respectively. We summarize our results in Table 1.

1.2 Related work

The first results obtaining generalization bounds via algorithmic stability can be found in (Rogers and Wagner, 1978; Devroye and Wagner, 1979). The work of Bousquet and Elisseeff (2002) continued this direction and provided guarantees for general supervised learning algorithms with regularization. Hardt et al. (2016) showed algorithmic stability for stochastic gradient descent without explicit regularization by exploiting gradient descent's non expansivity and Mou et al. (2018) showed stability for the stochastic gradient Langevin algorithm. Several works have focused on obtaining algorithmic stability for smooth convex functions, where the dependence on number of iterations for the optimization rates is faster than for the class of Lipschitz convex functions. In particular, Chen et al. (2018) showed that accelerated gradient descent (Nesterov, 1983) enjoys certain uniform stability for quadratic functions, pointed out the trade-off between stability and optimization showing as a result a lower bound on the optimization rates for this algorithm of $\widetilde{\Omega}(1/T^2)$. Attia and Koren (2021) showed that for general convex smooth functions the uniform stability grows exponentially with the number of iterations. Attia and Koren (2022) designed a method that obtains a stable algorithm from an optimization algorithm for Euclidean smooth convex functions by using the latter as a black-box. They also show stability for a specific unaccelerated algorithm designed for convex smooth functions with respect to $\|\cdot\|_p$, for $p \in [1,2]$. They pose an open problem, restated in the introduction, whose solution implies that it is possible to use optimization methods for convex and smooth functions with ℓ_p-norm regularity to achieve statistical guarantees. In this work, we solve the open problem in an affirmative with the additional benefit of recovering *optimal* statistical guarantees. Zhang et al. (2022) use Euclidean regularization in order to obtain a procedure that uses an optimization algorithm as a black-box and obtains a differentially private algorithm.

Concerning optimization, (Nemirovskii and Nesterov, 1985; Khachiyan et al., 1993) presented modifications of Nesterov's accelerated gradient descent in order to obtain accelerated algorithms for convex and smooth function with respect to p-norms. An algorithm and proof can also be found in (d'Aspremont et al., 2018, Section 4). Also for non-Euclidean regularity assumptions, Diakonikolas and Guzmán (2024) provided algorithms for objectives consisting of the sum of a smooth term and a proximable term, with a special focus on achieving acceleration when the proximable term is uniformly convex. The studied stability for $p \in [1,2]$ with their specific algorithm via regularization, as opposed to our black-box algorithm transformation. Juditsky and Nesterov (2014) studied optimization algorithms for Lipschitz, uniformly convex problems with respect to non-Euclidean norms.

We require for the algorithms that we transform that they

	LB (Non-Euclidean, ℓ_p-norm)	UB (Euclidean ℓ_2-norm)	UB (Non-Euclidean, ℓ_p-norm)
$d \le n$	$\widetilde{\Omega}_p(LR^2 \frac{d^{1/2-1/\hat{p}}}{n^{1/2}})$ (Levy and Duchi, 2019)	$\widetilde{\mathcal{O}}_p(LR^2 \left(\frac{1}{n}\right)^{1/2})$ (Attia and Koren, 2022)	$\widetilde{\mathcal{O}}_p(LR^2 \frac{d^{1/2-1/\hat{p}}}{n^{1/2}})$ (Corollary 10)
$d > n$	$\widetilde{\Omega}_p(LR^2 \left(\frac{1}{n}\right)^{1/\hat{p}})$ (Theorem 7)		$\widetilde{\mathcal{O}}_p(LR^2 \left(\frac{1}{n}\right)^{1/\hat{p}})$ (Corollary 11)

Table 1: Excess risk bounds for black-box reduction algorithms for ERM with loss functions in $\mathbb{R}^d$ that are L-smooth over the ball of radius R w.r.t. ℓ_p-norm for $n = \Omega_p(1)$. We denote $\hat{p} = \max\{p, 2\}$. The lower bound on the excess risk (LB) is for estimators within an ℓ_p-ball of radius R and by observing that the Lipschitz constant is bounded by $2LR$. The lower bound that we establish in the case $d > n$ improves upon the lower bound derived in (Sridharan, 2012) as discussed in Section 1.2. The upper bounds on the excess risk (UB) are achieved by the black-box reduction USOL2 (Attia and Koren, 2022) when $p = 2$ and USOLP (Algorithm 1) for $p \ge 1$.

either work with a compact convex set as a constraint or that their iterates are guaranteed to naturally stay in one such set. A variety of unconstrained algorithms for convex problems have the property that their iterates remain in a ball of center a minimizer x^* and radius a constant times the initial distance to x^*. For instance, this is satisfied by gradient descent for convex smooth functions with respect to $\|\cdot\|_2$. Similarly, it is possible to show that the iterates of Nesterov's accelerated gradient descent, as well as their more general accelerated hybrid proximal extragradient algorithms, remain bounded in a ball around a minimizer of radius $O(R_0)$, where R_0 is the initial distance to this minimizer (Monteiro and Svaiter, 2013, Theorem 3.10). The iterates x_t of mirror descent algorithms with a (μ, ν)-uniformly convex regularizer ψ with respect to a norm $\|\cdot\|$, satisfy $\frac{\mu}{\nu}\|x^* - x_t\|^\nu \le D_\psi(x^*, x_t) \le D_\psi(x^*, x_0)$.

There are known lower bounds for the minimax risk for a class of random loss functions in $\mathcal{B}_{\|\cdot\|_p}(1)$ which are linear and 1-Lipschitz w.r.t. ℓ_p-norm. For the low-dimensional case, i.e., $d \le n$, and $p \ge 1$, Levy and Duchi (2019) provided lower bounds for the minimax risk which are $\widetilde{\Omega}_p(d^{1/2-1/\hat{p}}/n^{1/2})$ where $\hat{p} = \max\{p, 2\}$ and they showed that a mirror descent with a specific distance function achieves these rates. In the high-dimensional case when $d > n$, the work of Sridharan (2012, Section 8) combined with the fact that the ℓ_{p^*} space cannot have Rademacher type greater than $p^* = p/(p-1)$ when $p^* \in [1, 2]$ (Albiac and Kalton, 2006, Remark 6.2.11.g), implies that the lowest minimax risk for a class of random linear 1-Lipschit losses in $\mathcal{B}_{\|\cdot\|_p}(1)$ cannot be less or equal to $c_p n^{-1/p-\delta}$ when $p \ge 2$ for any universal constants $c_p, \delta > 0$ where c_p may depend on p. We show a slightly stronger of $\Omega_p(1/n^{1/\hat{p}})$ for $\hat{p} = \max\{p, 2\}$ in Theorem 7 using a very different technique, inspired to the construction given by Levy and Duchi (2019). We match these rates with our black-box reduction algorithm as shown in Theorem 8 for $p \in [1, 2]$.

Notation and terminology. We denote $g^*(y) = \sup_{x \in \mathbb{R}^d}\{\langle x, y\rangle - g(x)\}$ the Fenchel dual function of a function g. We denote a general norm as $\|\cdot\|$ and its dual norm as $\|\cdot\|_*$. The ℓ_p-norm is denoted as $\|\cdot\|_p$. We use $\mathcal{B}_{\|\cdot\|_p}(x, R)$ for the ℓ_p ball of center x and radius R, and we use $\mathcal{B}_{\|\cdot\|_p}(R)$ if it is centered at 0. As it is usual, we say $\hat{x}$ is an ε-minimizer of the problem $\min_{x \in \mathcal{X}} f(x)$ if $f(\hat{x}) - \min_{x \in \mathcal{X}} f(x) \le \varepsilon$.

2 PRELIMINARIES

2.1 Generalizations of Smoothness and Convexity

In this work we are interested in the optimization of convex functions that are smooth with respect to non-Euclidean geometries, over compact convex sets $\mathcal{X} \subseteq \mathbb{R}^d$, i.e., when measuring distances with p-norms. The standard notions of strong convexity and smoothness are not well suited for this setting. For example, function $\psi(x) = \frac{1}{2}\|x\|_p^2$ has its strong convexity parameter bounded above by $1/d^{1-\frac{2}{p}}$ (Diakonikolas and Guzmán, 2024) which is small for high-dimensional problems and yields dimension-dependent rates when $p \gg 2$. The same is true for any strongly convex function w.r.t $\|\cdot\|_p$ whose range is bounded above by a constant on a unit ℓ_p-ball (d'Aspremont et al., 2018, Example 5.1).

In order to get dimension-independent rates, one can exploit instead the regularity of the objective function in terms of uniform convexity and Hölder smoothness, which are generalizations of strong convexity and smoothness.

Definition 1 (Uniform convexity). *We say f is (μ, ν)-uniformly convex with respect to norm $\|\cdot\|$ if for any $x, y \in \mathbb{R}^d$ we have*

$$f\left(tx + (1-t)y\right) \ge tf(x) + (1-t)f(y) + t(1-t)\frac{\mu}{\nu}\|x-y\|^\nu,$$

or, equivalently, for differentiable f:

$$f(y) \ge f(x) - \langle \nabla f(x), x - y\rangle + \frac{\mu}{\nu}\|x-y\|^\nu, \forall x, y \in \mathbb{R}^d,$$

which is shown, for example in (Nesterov, 2018, Section 4.2.2).

Note that the above definition simplifies to the notion of strong convexity when $\nu = 2$.

Definition 2 (Hölder smoothness). *We say that $f : \mathbb{R}^d \to \mathbb{R}$ is (L, ν)-Hölder smooth w.r.t. $\|\cdot\|$, if there exists $L > 0$ and $\nu \in [1, 2]$, such that for all $x, y \in \mathbb{R}^d$*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{\nu}\|x - y\|^\nu,$$

or equivalently, when f is also convex,

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\|^{\nu-1} \text{ for all } x, y \in \mathbb{R}^d.$$

We say that a function is L-smooth if it is $(L, 2)$-Hölder smooth.

The notion of uniform convexity allows us to capture function regularity in terms of non-Euclidean ℓ_p-norm geometry without needing to resort to dimension-dependent bounds. The function of ℓ_p-norm squared $\psi(x) = \frac{1}{2}\|x\|_p^2$ has its strong convexity constant bounded above by $1/d^{1-2/p}$, for $p \geq 2$ (Diakonikolas and Guzmán, 2024), which becomes small in high-dimensional settings. In fact, d'Aspremont et al. (2018, Example 5.1) show that *any* strongly convex function w.r.t. $\|\cdot\|_p$ with a constant bounded range in a unit ℓ_p-ball must have its strong convexity smaller than $1/d^{1-2/p}$ for $p \geq 2$. The notion of uniform convexity (Definition 1) allows to show that $\frac{1}{p}\|x\|_p^p$ is $(2^{2-p}, p)$-uniformly convex w.r.t $\|\cdot\|_p$, i.e., its uniform convexity constant is *dimension-independent*, see Lemma 15 and references therein. Ball et al. (1994) show that the bound on the uniform convexity is no longer dimension-dependent with the aforementioned $\psi(x) = \frac{1}{p}\|x\|_p^p$ being a specific example of a function whose uniform convexity constant is dimension-independent.

The fact that the uniform convexity constant of $\psi(x) = \frac{1}{p}\|x\|_p^p$ w.r.t. $\|\cdot\|_p$ norm is dimension-independent combined with optimization algorithms for uniformly convex objectives with dimension-independent convergence rates, e.g., Generalized AGD+ (Diakonikolas and Guzmán, 2024), allows us to derive learning algorithms with both, dimension-independent stability and optimization rates.

Lemma 3 (Duality of Hölder smoothness and uniform convexity). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be (μ, ν)-uniformly convex function w.r.t. $\|\cdot\|$. Then its Fenchel dual f^* is $((\frac{\nu^*}{2\mu(\nu^*-1)})^{\nu^*-1}, \nu^*)$ - Hölder smooth w.r.t. $\|\cdot\|_*$, where $\nu^* = \nu/(\nu - 1)$.*

The proof is given in Section 7.1 and follows a similar reasoning as the proof of (Shalev-Shwartz, 2007, Lemma 15) which shows duality between strongly convex and smooth functions. A different proof can be found also in (Zalinescu, 2002, Theorem 3.5.5).

2.2 Algorithmic stability

There are different notions of algorithmic stability that are useful for controlling generalization error (Bousquet and Elisseeff, 2002; Kutin and Niyogi, 2002). Herein, we use the notion of *uniform algorithmic stability* introduced by Bousquet and Elisseeff (2002).

Definition 4 (Uniform stability). *An algorithm $\mathcal{A}$, which outputs an estimator $\hat{x}$ and $\hat{x}'$ when given the datasets S and S' respectively, that differ in at most one sample, is said to be $\mathcal{E}_{\text{stab}}$-*uniformly stable*, if*

$$\sup_{z \in \mathcal{Z}} |\ell(\hat{x}; z) - \ell(\hat{x}'; z)| \leq \mathcal{E}_{\text{stab}}(\mathcal{A}).$$

A known result of Bousquet and Elisseeff (2002) is that an estimator $\hat{x}$ obtained from a uniformly stable algorithm $\mathcal{A}$ has bounded the generalization error term in (1) as

$$\mathbb{E}_S\left[f(\hat{x}) - f_S(\hat{x})\right] \leq \mathcal{E}_{\text{stab}}(\mathcal{A}). \tag{2}$$

3 STABILITY AFTER UNIFORMLY CONVEX REGULARIZATION

We study the uniform stability of estimators that approximately solve the regularized problem

$$x_\mu^* \in \arg\min_{x \in \mathcal{X}} f_S^{(\mu)}(x) \stackrel{\text{def}}{=} f_S(x) + \mu\psi(x), \tag{3}$$

where $\psi(x) : x \in \mathbb{R}^d \to \mathbb{R}$ is a $(1, \nu)$-uniformly convex regularizer that attains its minimum at a point $x_0 \in \mathbb{R}^d$. For example, a viable choice for a $(1, p)$-uniformly convex regularizer based on the ℓ_p-norm, which is also used in Section 4, is $\psi(x) = \frac{2^{p-2}}{p}\|x - x_0\|_p^p$ for $p \geq 2$ and some $x_0 \in \mathcal{X}$, cf. Lemma 15. Our results assume an upper bound on the regularizer distance between the minimum of the ERM and the regularizer, i.e., $\psi(x^*) - \psi(x_0) \leq D^\nu$ for $x^* \in \arg\min_{x \in \mathcal{X}} f_S(x)$, for $D \geq 0$.

We show that for the right choice of μ, the approximate solutions $\hat{x}$ to the optimization problem in (3) are uniformly stable while also having bounded distance to the original solution of the *unregularized* empirical risk.

Lemma 5 (Stability after uniformly convex regularization). *Let $\nu \geq 2$, $\mu > 0$, and $\mathcal{A}$ be an algorithm that for a dataset S returns an $\hat{\varepsilon}$-minimizer $\hat{x}$ of the regularized ERM $f_S^{(\mu)}$, where the losses $\ell(\cdot, z)$ are G-Lipschitz and the added regularizer $\mu\psi(x)$ is (μ, ν)-uniformly convex w.r.t. $\|\cdot\|$. If the accuracy is bounded as $\hat{\varepsilon} \leq \min\{\mu D^\nu, (\frac{\nu}{\mu})^{1/(\nu-1)}(\frac{2G}{n})^{\nu/(\nu-1)}\}$, then the algorithm $\mathcal{A}$ is uniformly stable as*

$$\mathcal{E}_{\text{stab}}(\mathcal{A}) \leq 3\left(\frac{2\nu}{n\mu}G^\nu\right)^{1/(\nu-1)},$$

while the optimization error of $\hat{x} = \mathcal{A}(S)$ on the unregularized empirical risk is upper bounded as

$$\mathcal{E}_{\text{opt}} \stackrel{\text{def}}{=} f_S(\hat{x}) - \min_{x \in \mathcal{X}} f_S(x) \leq 2\mu D^\nu.$$

The proof is given in Section 8.1 and follows a similar technique as the proof of (Attia and Koren, 2022, Lemma 7), but differs by using properties of uniformly instead of strongly convex regularizers and by considering only approximate minimizers of the regularized ERM.

Recall that the expected excess risk is bounded by the sum of the optimization error and the stability error. By (2), which is due to Bousquet and Elisseeff (2002), we have that stability controls the generalization error. In the result that follows, we find the optimal choice of μ and the necessary precision for the optimization error on $f_S^{(\mu)}$, given n, ν, G, and D, for which the upper bound on the expected excess risk is minimized.

Theorem 6 (Excess risk bound after uniformly convex regularization)**.** *Let $\nu \geq 2$ and $\hat{x}_\mu$ be a $(6GD/n^{1+1/\nu})$-minimizer of the regularized ERM $f_S^{(\mu)}(x)$ in (3) with G-Lipschitz losses $\ell(\cdot, z)$ w.r.t. $\|\cdot\|$, for a constant $\mu = \mathcal{O}_\nu(D^{1-\nu}G/n^{1/\nu})$. Then the expected excess risk of $\hat{x}_\mu$ is upper bounded as*

$$\mathbb{E}_S\left[\delta f(\hat{x}_\mu)\right] \leq 8GD\left(\frac{1}{n}\right)^{1/\nu}.$$

The proof is in Section 8.2 and consists of bounding the generalization and optimization errors using the uniform stability bound derived in Lemma 5, finding the optimal choice of μ that minimizes the excess risk, and ensuring that the error of the estimate $\hat{x}_\mu$ is sufficiently small.

Note that the results of Lemma 5 and Theorem 6 study properties of an approximate minimizer of the regularized ERM and are not concerned by how the approximate minimizer is computationally obtained. In cases when the bound on the regularizer implies a bounded domain containing the global minimizer, e.g., when $\mathcal{X} = \mathcal{B}_{\|\cdot\|}(x_0, R)$ for $R > 0$ and $x^* \in \mathcal{X}$, the smoothness of f implies its Lipschitzness

$$\|\nabla f(x)\| = \|\nabla f(x) - \nabla f(x^*)\| \leq L\|x - x^*\| \leq 2LR, \quad (4)$$

where we used that $\nabla f(x^*) = 0$. As a result, Lemma 5 and Theorem 6 apply also when the loss function has only prescribed smoothness inside of a bounded domain containing the global minimizer since $G \leq 2LR$.

3.1 Lower bound on the expected excess risk

In the following theorem we extend the technique of (Levy and Duchi, 2019, Proposition 1) to the high-dimensional case using sparse distributions and establish a lower bound on the expected excess risk for the specific case when the regularity of the loss $\ell(\cdot, z)$ is in respect to the ℓ_p-norm and $x^* \in \mathcal{B}_{\|\cdot\|_p}(R)$ for $R > 0$.

Theorem 7. *For $p \geq 1$, $48n < d$, and $\ell(x; z)$ that is linear and G-Lipschitz w.r.t. $\|\cdot\|_p$, we have that the expected excess risk in $\mathcal{B}_{\|\cdot\|_p}(R)$ is lower bounded as*

$$\inf_{\hat{x} \in \mathcal{B}_{\|\cdot\|_p}(R)} \sup_{P \in \mathcal{P}} \mathbb{E}_{S\sim P^n}\left[\delta f(\hat{x})\right] \geq \frac{GR}{12} \min\left\{1, \sqrt{\frac{\log(d)}{n}}\right\}$$

when $1 \leq p \leq 1 + 1/\log(d)$ and

$$\inf_{\hat{x} \in \mathcal{B}_{\|\cdot\|_p}(R)} \sup_{P \in \mathcal{P}} \mathbb{E}_{S\sim P^n}\left[\delta f(\hat{x})\right] \geq \frac{GR}{16}\left(\frac{1}{n}\right)^{1/\hat{p}}$$

when $p > 1 + 1/\log(d)$ where $\hat{p} = \max\{p, 2\}$.

The proof can be found in Section 8.3. The theorem shows that there exists a family of probability distributions and losses for which the performance of the best estimator $\hat{x} \in \mathcal{B}_{\|\cdot\|_p}(R)$ in terms of the excess risk cannot be lower than a certain bound. This result improves over the one implied by Sridharan (2012, Section 8) as explained in Section 1.2. From (Levy and Duchi, 2019, Proposition 1), we have that the risk is lower bounded as $\widetilde{\Omega}(GRn^{-1/2})$ when $d \leq 48n$, where G and R are w.r.t. the ℓ_p-norm.

Since any linear loss function in Theorem 7 is also convex, the lower bound also applies to problems defined for convex G-Lipschitz losses, denoted as $\mathcal{G}(G)$:

$$\frac{1}{16}GR\left(\frac{1}{n}\right)^{1/\hat{p}} \leq \inf_{\hat{x} \in \mathcal{B}_{\|\cdot\|_p}(R)} \sup_{\substack{P \in \mathcal{P} \\ \ell(\cdot,z) \in \mathcal{G}(G)}} \mathbb{E}_{S\sim P^n}\left[\delta f(\hat{x})\right].$$

By Theorem 6, we obtain a matching upper bound on the excess risk when $p \geq 2$ using $\psi(x) = (2^{p-2}/p)\|x\|_p^p$ as the regularizer, which is $(1, p)$-uniformly convex by Lemma 15. From the lemma, we get that the excess risk of $\hat{x}$ that is a $(12GR/n^{1+\frac{1}{p}})$-minimizer and satisfies

$$\sup_{\substack{P \in \mathcal{P} \\ \ell(\cdot,z) \in \mathcal{G}(G)}} \mathbb{E}_{S\sim P^n}\left[\delta f(\hat{x})\right] \leq 16GR\left(\frac{1}{n}\right)^{1/p},$$

where R denotes the radius of the domain $\mathcal{B}_{\|\cdot\|_p}(x_0, R)$ containing x^*, which comes from the form of $\psi(x)$ and its bounded range as $D = 2^{1-1/p}p^{-1/p}R \leq 2R$.

As a consequence, the estimator $\hat{x}$ gives an optimal bound up a constant in p on the excess risk for ERM with G-Lipschitz convex losses

$$\frac{1}{16}GR\left(\frac{1}{n}\right)^{1/p} \leq \inf_{\hat{x} \in \mathcal{B}_{\|\cdot\|_p}(R)} \sup_{\substack{P \in \mathcal{P} \\ \ell(\cdot,z) \in \mathcal{G}(G)}} \mathbb{E}_{S\sim P^n}\left[\delta f(\hat{x})\right] \leq 16GR\left(\frac{1}{n}\right)^{1/p},$$

when $d > 48n$. The optimality of the rate does not depend on the smoothness of $\ell(\cdot, z)$ since the upper bound from Theorem 6 applies regardless of how smooth the loss is and the lower bound applies for linear losses that are 0-smooth.

4 NON-EUCLIDEAN BLACK-BOX UNIFORM STABILITY

We show that applying an optimization algorithm to the ERM problem with the added regularization term based on $\|\cdot\|_p^p$ combined with a restart scheme yields a black-box reduction to an *optimally* stable learning algorithm with the same convergence rate on the optimization error. Let

$$x_\mu^* \in \arg\min_{x\in\mathcal{X}} f_S^{(\mu)}(x) \stackrel{\text{def}}{=} f_S(x) + \mu\frac{\alpha}{p}\|x-x_0\|_p^p, \quad (5)$$

where $\mu > 0$ and

$$\alpha = \begin{cases} 2^{p-2} & \text{for } p \geq 2 \\ 2^{2p-3}\left(1-\frac{1}{p}\right)^{p-1} & \text{for } p \in (1,2) \end{cases}$$

ensures that $\psi(x) = \frac{\alpha}{p}\|x-x_0\|_p^p$ is $(1,p)$-uniformly convex for $p \geq 2$ and $(1,p)$-Hölder smooth for $p \in (1,2)$ w.r.t. ℓ_p-norm, see Lemma 15. We will also use that $\psi(x)$ is *locally* smooth for $p \geq 2$ and strongly convex for $p \in (1,2)$ w.r.t. ℓ_p-norm, see Lemma 14.

The black-box reduction algorithm USOLP, given in Algorithm 1, transforms an optimization algorithm $\mathcal{A}$ meant for convex Hölder smooth objective functions to an optimally stable algorithm by solving (5) with an appropriate chosen $\mu > 0$. Adding the regularization term $\psi(x)$ comes with two benefits: (i) statistically, it guarantees the minimizer is uniformly stable due to Lemma 5, and (ii) in optimization, it makes the objective *uniformly convex* and smooth when $p \geq 2$ or convex and *Hölder smooth* when $p \in (1,2)$, which with an appropriate restarting scheme yields appropriate convergence rate that in some situations is faster compared to the original algorithm $\mathcal{A}$.

Let $\mathcal{A}(f_S^{(\mu)}, x_0, R, \hat{\varepsilon})$ be an optimization algorithm for convex Hölder smooth objective functions that takes an initial point x_0, an upper bound $R \geq 0$ on the distance of x_0 to the minimizer, and a required target accuracy $\hat{\varepsilon}$. The algorithm then outputs a point $\hat{x} \in \mathcal{B}_{\|\cdot\|_p}(x_0, R)$ such that

$$f_S^{(\mu)}(\hat{x}) - \inf_{x\in\mathcal{B}_{\|\cdot\|_p}(x_0,R)} f_S^{(\mu)}(x) \leq \hat{\varepsilon}, \quad (6)$$

where to achieve the goal the algorithm performs at most $\hat{T}$ gradient oracle calls.

The assumption, that the optimization algorithm outputs a point in $\mathcal{B}_{\|\cdot\|_p}(x_0, R)$ is formalized below and is used in Lemma 14 to control smoothness and strong convexity of $\|\cdot\|_p^p$ when $p \geq 2$ and $p \in (1,2)$ respectively.

Assumption 1. *For any initial point x_0, and for some $R > 0$, the iterates and the output of Algorithm $\mathcal{A}$ optimizing f with a minimizer x^* are in the domain $\mathcal{X} = \mathcal{B}_{\|\cdot\|_p}(x_0, R)$ containing x^*.*

We note that the two main cases where this assumption is satisfied is for (i) algorithms that work when constrained to $\mathcal{X}$, or for (ii) unconstrained algorithms whose iterates naturally stay in a ball around the minimizer $\mathcal{B}_{\|\cdot\|_p}(x^*, 2R)$, for $R = \mathcal{O}(\|x_0 - x^*\|)$. As explained in Section 1.2, many known algorithms fall in the second category. Note that Assumption 1 guarantees also the condition on the bounded range $\psi(x^*) - \psi(x_0) \leq D^p$ in Lemma 5, which for $\psi(x) = (\alpha/p)\|x - x_0\|_p^p$ translates into the requirement that $x^* \in \mathcal{B}_{\|\cdot\|_p}(x_0, R)$ with $R = (p/\alpha)^{1/p}D$.

Algorithm 1 USOLP($\mathcal{A}$) for ℓ_p-structured learning

Input: Algorithm $\mathcal{A}$, estimate R, $x_0 \in \mathbb{R}^d$, $T = \widetilde{\Omega}_p(n^{1/\gamma})$, $p \geq 1$
$\hat{p}_1 = \min\{p, 2\}$
$\hat{p}_2 = \max\{p, 2\}$
if $p \geq 2$ **then**
  $\mu = (4C/T)^\gamma p\alpha^{2/p-1}LR^{2-p}$
  $r = \lceil \log_2((T/C)^\gamma \alpha^{-2/p})\rceil$
else if $p \in (1,2)$ **then**
  $\mu = (C/T)^\gamma \frac{4}{p-1}\alpha^{-2/p}LR^{2-p}$
  $r = \lceil \log_2(R^{p-2}(T/C)^\gamma/(2\alpha))\rceil$
end if
$R_0 = R$
$\hat{\varepsilon}_0 = LR_0^{\hat{p}_1}/\hat{p}_1$
for $i = 1, \ldots, r$ **do**
  $\hat{\varepsilon}_i = \hat{\varepsilon}_{i-1}/2$
  $x_i = \mathcal{A}(f_S^{(\mu)}, x_{i-1}, R_{i-1}, \hat{\varepsilon}_i)$
  $R_i = \left(\frac{\hat{p}_2 R^{2-\hat{p}_1}\hat{\varepsilon}_i}{\mu(\hat{p}_1-1)}\right)^{1/\hat{p}_2}$
end for
Output: $x_T \stackrel{\text{def}}{=} x_r$

Since we study the stability of first order methods for smooth objectives whose convergence rates are based on the smoothness of the objective function, in this section we pose the results for L-smooth loss functions using the fact that $G \leq 2LR$ on a bounded domain containing the minimizer as given in (4).

Theorem 8 (Black-box uniform stability)**.** *Let $p > 1$, the loss function $\ell : \mathbb{R}^d \to \mathbb{R}$ be convex and L-smooth w.r.t. $\|\cdot\|_p$ and $\mathcal{A}$ be an optimization algorithm satisfying Assumption 1 with $R > 0$ that has a convergence rate $C\hat{L}\|x_0 - x^*\|_p^{\hat{p}_1}/\hat{T}^\gamma$ for convex, $(\hat{L}, \hat{p}_1)$-Hölder smooth objective functions, where $\hat{p}_1 = \min\{p, 2\}$. Then, for $T = \widetilde{\Omega}_p(n^{1/\gamma})$, the iterate x_T produced by* USOLP($\mathcal{A}, T$) *initialized at x_0, satisfies,*

1. $\mathcal{E}_{\text{stab}}(\text{USOLP}(\mathcal{A}, T)) = \widetilde{\mathcal{O}}_p((T^\gamma/n)^{\frac{1}{\hat{p}_2-1}}LR^2)$,
2. $f_S(x_T) - f_S(x^*) = \widetilde{\mathcal{O}}_p(LR^2/T^\gamma)$,

for $p \in (1,\infty)$ where $\hat{p}_2 = \max\{p, 2\}$ and $T = \sum_{i=1}^r \hat{T}_i$ is the sum of gradient oracle calls from all stages with each stage taking $\hat{T}_i$ gradient oracle calls.

The proof, which we provide in Section 9.1, consists of two parts. Firstly, we use a restarting scheme to transfer the algorithm $\mathcal{A}$, which has convergence rate $\hat{\varepsilon} = \mathcal{O}(\hat{L}\hat{T}^{-\gamma})$ when applied to the minimization of a convex $(\hat{L}, 2)$-Hölder smooth objective function, to an algorithm with a potentially *faster* convergence rate of $\hat{\varepsilon} = \mathcal{O}(\hat{L}\mu^{-\frac{2}{p-2}}T^{-\gamma\frac{p-2}{p}})$ when applied to the minimization of a (μ, p)-*uniformly convex* $(\hat{L}, 2)$-Hölder smooth objective, where $\hat{\varepsilon}$ is the wanted accuracy of the algorithm. The second part consists of choosing the regularization parameter μ controlling the uniform convexity of the objective on the order of $\mu = \mathcal{O}(\hat{\varepsilon})$, which is the prescribed condition on the required accuracy in Lemma 5. This choice of μ brings the convergence rate of the optimization error to the original rate $\hat{\varepsilon} = \mathcal{O}(\hat{L}T^{-\gamma})$ but with the added benefit of being uniformly stable due to the choice of μ. Similar argument can be made when $p \in (1, 2)$ but with the exception of having a $(\hat{L}, p)$-Hölder smooth and (μ, p)-uniformly convex, i.e. μ-strongly convex objective.

The reduction nearly preserves the optimization convergence rate in the sense that $\widetilde{O}_p(LR^2/T^\gamma)$ is equal up to a constant factor depending on p and a log factor, to the convergence rate $\mathcal{O}_p(LR^2/\hat{T}^\gamma)$ of the optimization algorithm when applied to a smooth convex objective function. Here T denotes the cumulative sum of the number of gradient calls in all r restarting stages and $\hat{T}$ is the number of gradient oracle calls of the optimization algorithm.

Remark 9 (The case of $p \leq 1 + \log^{-1}(d)$)**.** *The results in Theorem 8 hold also in the case when $p \leq 1 + \log^{-1}(d)$ up to $\log(d)$ factors. This is a consequence of $\|x\|_1 = \Theta(\|x\|_{\hat{p}})$ for $\hat{p} = 1 + \log^{-1}(d)$, so an L-smooth function w.r.t. $\|\cdot\|_1$ is an $\mathcal{O}(L)$-smooth function w.r.t. $\|\cdot\|_{\hat{p}}$. We can then apply the results of Theorem 8 with $\|\cdot\|_{\hat{p}}$ and the constants depending on $\hat{p}$ will be modified in the bounds by $\mathcal{O}(\log(d))$ factors.*

The rates depending on γ will differ based on the optimization algorithm we use. For $p \geq 2$, we have optimization algorithms, such as GeneralizedAGD+ (Diakonikolas and Guzmán, 2024), whose rates are in the range $\gamma \in (0, 1 + 2/p]$, and employing USOLP reduction on an algorithm with the optimal rate $\gamma = 1 + 2/p$ yields a learning algorithm that is $\widetilde{\mathcal{O}}_p(T^{1+2/p})$-uniformly stable. When $p \in [1, 2)$ the convergence rates are in the range $\gamma \in (0, 3p/2 - 1]$ where the optimal upper limit translates into in $\widetilde{\mathcal{O}}(T^{1+p^*/2})$-uniformly stable algorithms for $1/p^* + 1/p = 1$. The upper limits for both regimes, $p \in [1, 2)$ and $p \geq 2$, are attained by (Diakonikolas and Guzmán, 2024; d'Aspremont et al., 2018) and the proof of optimality of the rates can be found in (Guzmán and Nemirovski, 2015).

The stability rates for the right choice of p and μ yield min-max rates that achieve an optimal upper bound for the excess risk up to a constant factor in p. In Corollary 10 we show that using USOLP with $p = 2$ achieves an excess risk upper bounded on the order of $d^{1/2-1/p}n^{-1/2}$, which is the same as that achieved by the Euclidean black-box reduction algorithm (Attia and Koren, 2022, Theorem 6) when we choose the number of iterations such that the upper bound on the excess risk is minimized.

Corollary 10 (Excess risk bounds of USOLP for $d \leq n$)**.** *Under the assumptions of Theorem 8, we have that USOLP applied to the regularized risk minimization in (3) with $\psi(x) = \frac{1}{2}\|x - x_0\|_2^2$ and μ chosen to be optimal w.r.t. smallest excess risk bound as in Theorem 6, produces an estimate x_T for which the excess risk is upper bounded as*

$$\mathbb{E}_S\left[\delta f(x_T)\right] = \mathcal{O}_p\left(LR^2\frac{d^{1/2-1/\hat{p}}}{n^{1/2}}\right),$$

where $\hat{p} = \max\{2, p\}$ when the total number of gradient oracle calls is on the order of $T = \widetilde{\Omega}_p((d^{\frac{1}{2}-\frac{1}{\hat{p}}}n^{\frac{1}{2}})^{\frac{1}{\gamma}})$ and $n = \Omega(d^{1/\hat{p}-1/2})$.

We provide the proof in Section 9.2. The proof consists of applying USOLP with the regularizer $\psi(x) = (1/2)\|x - x_0\|_2^2$, which by Theorem 6 achieves asymptotic sample complexity on the order of $n^{-1/2}$. The bound between ℓ_p-norms $\|x\|_2 \leq d^{1/2-1/\hat{p}}\|x\|_p$ where $\hat{p} = \max\{p, 2\}$ allows to express the regularity w.r.t. ℓ_p-norm at the cost of an extra $d^{1/2-1/\hat{p}}$ factor. The bound is optimal in the *low-dimensional case when $d \leq n$*, but becomes vacuous in d when the dimension of the problem is large.

The following Corollary 11 shows that in the *high-dimensional case $n > d$*, using USOLP with p based on the ℓ_p-norm regularity of the function yields an excess risk bound that is *independent of the dimension*, but is on the order of $n^{-1/\hat{p}}$.

Corollary 11 (Excess risk bounds of USOLP for $d > n$)**.** *Under the assumptions of Theorem 8, we have that USOLP applied to the regularized risk minimization in (3), where μ is chosen as in Theorem 6, produces an estimate x_T for which the excess risk is upper bounded as*

$$\mathbb{E}_S\left[\delta f(x_T)\right] = \mathcal{O}_p\left(LR^2\left(\frac{1}{n}\right)^{1/\hat{p}_2}\right),$$

when the total number of gradient oracle calls is on the order of

$$T = \begin{cases} \widetilde{\Omega}_p(n^{\frac{3-p}{2}\frac{1}{\gamma}}) & \text{when } p \in (1, 2) \\ \widetilde{\Omega}_p(n^{(1-\frac{1}{p})\frac{1}{\gamma}}) & \text{when } p \geq 2 \end{cases}$$

and $n = \Omega_p(1)$.

The proof is in Section 9.3 and is based on applying Theorem 6 with the optimal choice of μ and ensuring that the required number of iterations to reach the necessary accuracy is performed. USOLP achieves optimal excess risk up

to a constant factor in p regardless of the convergence rate γ of the original algorithm, but for larger γ it will require fewer gradient oracle calls to reach it. The condition on n being large enough ensures that the combined smoothness of the regularized objective is on the same order as the smoothness of the unregularized empirical risk.

Our black-box reduction, when applied to an optimization algorithm with a suitably fast convergence rate, achieves the same rates on the excess risk, up to a logarithmic factor, as the optimal non-black-box algorithm *Generalized AGD+* (Diakonikolas and Guzmán, 2024) when applied to an ERM with uniform convex regularization. In the result that follows we show that the convergence rates of Generalized AGD+ in (Diakonikolas and Guzmán, 2024, Theorem 2) applied to an ERM with uniformly convex regularization combined with the stability results of Lemma 5 translate to uniform stability rates that are identical up to a logarithmic factor to the ones from our black-box reductions in Theorem 8 when $p \geq 2$ and we apply the reduction to an optimization algorithm with $\gamma = 1 + 2/p$, e.g. the accelerated first order algorithm in (d'Aspremont et al., 2018). This implies that USOLP achieves the best possible rates for the black-box reduction.

Proposition 12 (Uniform stability of the Generalized AGD+)**.** *Let $p \geq 2$. Then for a given T, the iterate x_T produced by Generalized AGD+ algorithm (Diakonikolas and Guzmán, 2024) applied to the minimization of the regularized empirical risk minimization in (3) with $\mu = \widetilde{\Omega}_p(T^{-(1+2/p)} R^{2-p} L)$ is uniformly stable as*

$$\mathcal{E}_{\text{stab}}(\mathcal{A}(T)) = \widetilde{O}_p\left(LR^2\left(\frac{T^{1+2/p}}{n}\right)^{1/(p-1)}\right)$$

and its optimization error on the empirical risk is upper bounded as

$$f_S(x_T) - f_S(x^*) = \widetilde{O}_p\left(LR^2\left(\frac{1}{T}\right)^{1+2/p}\right),$$

provided $T = \Omega_p(R^{2-p} n^{p/(p+2)})$.

The proof is given in Section 9.4. For $p = 2$ our rates simplify to the stability rates derived by (Chen et al., 2018; Attia and Koren, 2022) for the Euclidean case and extend them for $p > 2$. The rates match the rates of USOLP in Theorem 8 when $\gamma = p + 2/p$.

Corollary 13 (Excess risk bound of the Generalized AGD+)**.** *Let $p \geq 2$, $\ell : \mathbb{R}^d \to \mathbb{R}$ is convex L-smooth function w.r.t. $\|\cdot\|_p$, and μ is chosen as in Theorem 6. Then the Generalized AGD+ algorithm applied to the regularized risk minimization in (3) produces an estimate x_T for which the excess risk is upper bounded as*

$$\mathbb{E}_S\left[\delta f(x_T)\right] \leq 16LR^2\left(\frac{1}{n}\right)^{1/p},$$

when the number of iterations is on the order of $T = \widetilde{\Omega}_p(n^{\frac{p-1}{p+2}}) = \widetilde{\Omega}_p(n^{(1-\frac{1}{p})\frac{p}{p+2}})$.

The proof is given in Section 9.5 and is based on using the optimal choice of μ as given in Theorem 6 and ensuring that the Generalized AGD+ algorithm reaches sufficiently low error using its proved convergence rate from (Diakonikolas and Guzmán, 2024, Theorem 2). The upper bound matches the result in Corollary 11 for $p \geq 2$ when $\gamma = p + 2/p$.

5 CLASSIFICATION IN ℓ_p-BALLS

We describe an application to binary classification where the regularity of the data and the optimal estimator are given in non-Euclidean geometry. In this setup the generalization guarantees in Corollary 11 improve upon the ones given by posing Euclidean geometry regularity on the problem.

Let $p \geq 2$ and $p^* = p/(p-1)$. The classification task consists of n labeled points $(a_i, b_i) \in \mathbb{R}^d \times \{-1, 1\}$, $i \in [n]$ where $a_i \in \mathcal{B}_{\|\cdot\|_{p^*}}(R)$. Consider the generalized linear model $h(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i\in[n]} h_i(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i\in[n]} f(b_i\langle a_i, x\rangle)$ where $f : \mathbb{R} \to \mathbb{R}$ is convex, L-smooth, and twice differentiable, e.g., the logistic loss, the squared loss $f(t) = (1-t)^2$, or the smoothed hinge loss (Rennie and Srebro, 2005). Since $\nabla h_i(x) = f'(b_i\langle a_i, x\rangle) b_i a_i$ and the bound on Lipschitzness in (4), we have

$$\begin{aligned}\langle \nabla^2 h(x)v, v\rangle &= \frac{1}{n}\sum_{i=1}^n |f''(b_i\langle a_i, x\rangle)|\left(\sum_{j=1}^d v_j a_{i,j}\right)^2 \\ &\leq L\|v\|_p^2 \sum_{i=1}^n \frac{\|a_i\|_{p^*}^2}{n} \leq LR^2\|v\|_p^2,\end{aligned}$$

where we used the Hölder's inequality and the assumption on the data. As a result h is LR^2-smooth with respect to the $\|\cdot\|_p$ norm.

Since for $p^* \leq 2$, we have $\|x\|_2 \leq \|x\|_{p^*}$, the data a_i is also in a Euclidean ball of radius R and, as a result, the function is also (LR^2)-smoothness w.r.t. $\|\cdot\|_2$. The excess risk bound in the Euclidean case is $O(n^{-1/2})$, which is a factor of $n^{1/2-1/p}$ lower than the upper bound of $O(n^{1/p})$ derived in Theorem 6 using non-Euclidean regularization. However, depending on the problem setup, the distance to a minimizer measured with the Euclidean distance can be a factor $d^{\frac{1}{2}-\frac{1}{p}}$ greater than when measured in ℓ_p-norm, with the upper bound reached when the data is proportional to the vector of ones. Thus, in the high-dimensional case $d \gg n$, using the p-norm can lead to statistical bounds in Theorem 6 that are up to $(d/n)^{1/2-1/p}$ lower compared to using the Euclidean regularity.

In the case when $p \in (1, 2)$, the ℓ_p-norm regularized problem can have much higher smoothness that can have beneficial properties for the optimization algorithms while keep-

ing the same dependence on n. However, in this case, the distance to the minimizer measured in ℓ_p-norm can be a factor of the dimension greater than when measured in the Euclidean distance. Same as when $p \geq 2$, the improvement will depend on where the data and minimizer are located in terms of the ℓ_p-ball.

6 CONCLUSION

In this paper we studied uniform stability properties of first-order algorithms in non-Euclidean settings, i.e., for functions whose regularity is measured w.r.t. $\|\cdot\|_p$ norms. We developed new bounds for the uniform stability of approximate minimizers of empirical risk with uniformly convex regularization. The stability bounds, in combination with an observation that a restart scheme for minimization of uniformly convex objectives results into convergence rates that capture the uniform convexity of the objective, allowed us to design a black-box reduction scheme that gives an optimal trade-off between stability and convergence rate. We also showed that the black-box reduction yields optimal expected excess risk up to a constant factor in p. We provided a derivation of an excess risk lower bound for $p \geq 2$ in the high-dimensional setting.

Acknowledgements

Simon Vary and Patrick Rebeschini were funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number EP/Y028333/1]. David Martínez-Rubio was partially funded by the project IDEA-CM (TEC-2024/COM-89). For the purpose of Open Access, the authors have applied a CC BY public copyright licence to any Author Accepted Manuscript (AAM) version arising from this submission.

References

F. Albiac and N. J. Kalton. *Topics in Banach Space Theory*, volume 233. Springer, 2006.

A. Attia and T. Koren. Algorithmic instabilities of accelerated gradient descent. In *Advances in Neural Information Processing Systems*, volume 34, pages 1204–1214, 2021.

A. Attia and T. Koren. Uniform stability for first-order empirical risk minimization. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178, pages 3313–3332. PMLR, 2022.

K. Ball, E. A. Carlen, and E. H. Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Inequalities: Selecta of Elliott H. Lieb*, pages 171–190, 1994.

D. Baudry, N. Merlis, M. Benjamin Molina, H. Richard, and V. Perchet. Multi-armed bandits with guaranteed revenue per arm. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pages 379–387. PMLR, 2024.

O. Bousquet and A. Elisseeff. Stability and Generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Y. Chen, C. Jin, and B. Yu. Stability and Convergence Trade-off of Iterative Optimization Algorithms. *CoRR*, abs/1804.01619, 2018.

A. d'Aspremont, C. Guzmán, and M. Jaggi. Optimal affine-invariant smooth minimization algorithms. *SIAM Journal on Optimization*, 28(3):2384–2405, 2018.

L. Devroye and T. J. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Trans. Inf. Theory*, 25(5):601–604, 1979.

J. Diakonikolas and C. Guzmán. Complementary composite minimization, small gradients in general norms, and applications. *Mathematical Programming*, 208(1–2): 319–363, 2024.

J. Duchi. *Lecture Notes on Statistics and Information Theory*. 2023. URL *https://web.stanford.edu/class/stats311/lecture-notes.pdf*.

C. Guzmán and A. Nemirovski. On lower complexity bounds for large-scale smooth convex optimization. *Journal of Complexity*, 31(1):1–14, 2015.

M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1225–1234, 2016.

A. Juditsky and Y. Nesterov. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems*, 4(1):44–80, 2014.

L. Khachiyan, A. Nemirovski, and Y. Nesterov. Optimal methods for the solution of large-scale convex programming problems. *Modern Mathematical Methods in Optimization*, 1993.

S. Kutin and P. Niyogi. Almost-everywhere algorithmic stability and generalization error. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, page 275–282, 2002.

D. Levy and J. C. Duchi. Necessary and sufficient geometries for gradient methods. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

R. D. C. Monteiro and B. F. Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2):1092–1125, 2013.

W. Mou, L. Wang, X. Zhai, and K. Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical

viewpoints. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, pages 605–638, 2018.

A. S. Nemirovskii and Y. E. Nesterov. Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics*, 25(2):21–30, 1985.

Y. Nesterov. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. In *Doklady Akademii Nauk SSSR*, volume 269, pages 543–547, 1983.

Y. Nesterov. *Lectures on Convex Optimization*. Springer Optimization and Its Applications. 2018.

J. Renegar and B. Grimmer. A simple nearly optimal restart scheme for speeding up first-order methods. *Foundations of Computational Mathematics*, 22(1):211–256, 2022.

J. D. Rennie and N. Srebro. Loss functions for preference levels: Regression with discrete ordered labels. In *Proceedings of the IJCAI multidisciplinary workshop on advances in preference handling*, volume 1, 2005.

W. H. Rogers and T. J. Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, pages 506–514, 1978.

S. Shalev-Shwartz. *Online learning: theory, algorithms and applications*. PhD thesis, Hebrew University of Jerusalem, Israel, 2007.

S. Shalev-Shwartz, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

K. Sridharan. *Learning From An Optimization Viewpoint*. PhD thesis, Toyota Technological Institute at Chicago, 2012.

C. Zalinescu. *Convex analysis in general vector spaces*. 2002.

L. Zhang, K. K. Thekumparampil, S. Oh, and N. He. Bring your own algorithm for optimal differentially private stochastic minimax optimization. *Advances in Neural Information Processing Systems*, 35:35174–35187, 2022.

C. Zălinescu. On uniformly convex functions. *Journal of Mathematical Analysis and Applications*, 95(2):344–374, 1983.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Black-Box Uniform Stability for Non-Euclidean Empirical Risk Minimization: Supplementary Materials

7 SUPPORTING LEMMATA

Lemma 14 (Restricted smoothness and strong convexity of $\frac{1}{p}\|x\|_p^p$ in the p-ball)**.** *Let $\psi(x) = \frac{1}{p}\|x\|_p^p$ defined for $x \in \mathcal{B}_{\|\cdot\|_p}(R)$. Then $\psi(x)$ is*

1. *$(p-1)R^{p-2}$-smooth w.r.t ℓ_p-norm inside of $\mathcal{B}_{\|\cdot\|_p}(R)$ when $p \geq 2$, and*
2. *$(p-1)R^{p-2}$-strongly convex w.r.t ℓ_p-norm inside of $\mathcal{B}_{\|\cdot\|_p}(R)$ when $p \in (1,2)$.*

Proof. We have $\nabla^2\psi(x) = (p-1)\operatorname{diag}((|x_i|^{p-2})_{i=1}^d)$. When $p \geq 2$, we can bound

$$\langle \nabla^2\psi(x)v, v\rangle = (p-1)\sum_{i=1}^d v_i^2|x_i|^{p-2} \leq (p-1)\|v\|_p^2\|x\|_p^{p-2} \leq (p-1)R^{p-2}\|v\|_p^2,$$

where we used Hölder's inequality $\langle u, w\rangle \leq \|u\|_q\|w\|_{q^*}$, with $q = p/2$ and so $q^* = p/(p-2)$, and $\|x\|_p = R$ by $x \in \mathcal{B}_{\|\cdot\|_p}(R)$. Consequently, by Taylor approximation, we have for all $x \in \operatorname{conv}\{y, z\}$, where $y, z \in \mathcal{B}_{\|\cdot\|_p}(0, R)$, that

$$\psi(y) \leq \psi(z) + \langle \nabla\psi(z), y-z\rangle + \frac{1}{2}\langle \nabla^2\psi(x)(y-z), y-z\rangle \leq \psi(z) + \langle \nabla\psi(z), y-z\rangle + \frac{(p-1)R^{p-2}}{2}\|y-z\|_p^2.$$

When $p \in (1,2)$, we can bound

$$\begin{aligned}
\langle \nabla^2\psi(x)v, v\rangle &= (p-1)\sum_{i=1}^d v_i^2|x_i|^{2-p} \\
&\geq (p-1)\left(\sum_{i=1}^d |v_i|^{2/q}\right)^q \left(\sum_{i=1}^d |x_i|^{\frac{2-p}{1-q}}\right)^{1-q} \\
&= (p-1)\left(\sum_{i=1}^d |v_i|^p\right)^{2/p} \left(\sum_{i=1}^d |x_i|^p\right)^{\frac{p-2}{p}} \\
&= (p-1)\|v\|_p^2\|x\|_p^{p-2} \geq (p-1)\|v\|_p^2\left(\frac{1}{R}\right)^{2-p},
\end{aligned}$$

where the first line comes from the form of $\nabla^2\psi(x)$, the second line uses the reverse Hölder inequality for $q \geq 1$, in the third line we choose $q = 2/p \geq 1$, and the final line comes from $x \in \mathcal{B}_{\|\cdot\|_p}(R)$ and $p \in (1,2)$. □

Lemma 15. *For $p \geq 2$ and $x_0 \in \mathbb{R}^d$, the regularizer $\psi(x) \overset{\text{def}}{=} \frac{2^{p-2}}{p}\|x - x_0\|_p^p$ is $(1,p)$-uniformly convex with respect to $\|\cdot\|_p$ in $\mathbb{R}^d$. For $p \in (1,2)$, the regularizer $\psi(x) = (1/p)\|x - x_0\|_p^p$ is $(2^{3-2p}(p/(p-1))^{p-1}, p)$-Hölder smooth.*

The fact that $\psi(x)$ is uniformly convex is classical, see for instance (Zălinescu, 1983), that established $(2^{-\frac{p(p-2)}{p-1}}, p)$-uniform convexity of $\frac{1}{p}\|x\|_p^p$. In the following proof, we show the uniform convexity of ψ with a better constant.

Proof. For $p \geq 2$. Note that $\|x\|_p^p$ is a separable function, i.e., it has the form $\sum_{i=1}^d f_i(x_i)$. Thus, it is enough to show the uniform convexity of the one-dimensional case and add up all of the corresponding inequalities in order to obtain the result. In (Nesterov, 2018, Lemma 4.2.3), it is established that $\frac{2^{p-2}}{p}\|x\|_2^p$ is $(1,p)$-uniformly convex with respect to the Euclidean norm $\|\cdot\|_2$. Since in one dimension, all of the p-norms are the same, the 1-D result is proven, and the result follows.

For $p \in (1,2)$. Let $q = p/(p-1)$, we have that $q > 2$ and $(1/q)\|x\|_q^q$ is $(2^{2-q}, q)$-uniformly convex by the first part of the proof. By similar derivations as in (Boyd and Vandenberghe, 2004, Example 3.27) but with Hölder inequality, we get that the Fenchel dual of $(1/q)\|x\|_q^q$ is $(1/p)\|x\|_p^p$. From Lemma 3 combined with $(1/q)\|x\|_q^q$ being $(1,q)$-uniformly convex, we have that $(1/p)\|x\|_p^p$ is $(2^{3-2p}(p/(p-1)^{p-1}), p)$-Hölder smooth. □

Lemma 16 (Distance bound)**.** *Let $\psi(x)$ be an (μ,p)-uniformly convex regularizer of the empirical risk $f_S(x)$, i.e., $f_S^{(\mu)}(x) = f_S(x) + \psi(x)$, and define $x^* = \arg\min_{x\in\mathcal{X}} f_S(x)$, $x_\mu^* = \arg\min_{x\in\mathcal{X}} f_S^{(\mu)}(x)$, and $x_0 = \arg\min_{x\in\mathcal{X}} \psi(x)$. Then, if $\psi(x^*) - \psi(x_0) \leq D^p$ for $D > 0$, we also have that*

$$\mathrm{D}_\psi(x_\mu^*, x_0) \leq D^p,$$

where D_ψ is the Bregmann divergence of ψ

Proof. If we assume that we have $\psi(x^*) - \psi(x_0) \leq D^p$ for $D > 0$, we get

$$\begin{aligned}
\mathrm{D}_\psi(x_\mu^*, x_0) &= \psi(x_\mu^*) - \psi(x_0) - \langle \nabla\psi(x_0), x_0 - x_\mu^*\rangle \\
&\leq \psi(x_\mu^*) - \psi(x_0) \\
&\leq \psi(x_\mu^*) - \psi(x_0) + \frac{1}{\mu}\left(f_S(x_\mu^*) - f_S(x^*)\right) \\
&\leq \psi(x_\mu^*) - \psi(x_0) + \psi(x^*) - \psi(x_\mu^*) = \psi(x^*) - \psi(x_0) \leq D^p,
\end{aligned} \tag{7}$$

where the first inequality comes from x_0 being the minimum of ψ, the second inequality from x^* being minimum of f_S, the third inequality from x_μ^* being the minimum of $f_S^{(\mu)}$, and the last fourth inequality comes from the definition of D. □

The following is a generalization of (Attia and Koren, 2022, Lemma 1) for uniformly convex functions.

Lemma 17 (Upper bound on distance between perturbed minima)**.** *Let f_1 be convex and f_2 be (μ,p)-uniformly convex w.r.t. the norm $\|\cdot\|$. For $x_1 \in \arg\min_x f_1(x)$ and $x_2 \in \arg\min_x f_2(x)$ we have*

$$\|x_1 - x_2\| \leq \left(\frac{p}{\mu}\|\nabla f_1(x_1) - \nabla f_2(x_1)\|_*\right)^{\frac{1}{p-1}},$$

where $\|\cdot\|_*$ is the dual norm to $\|\cdot\|$.

Proof. By f_2 being (μ,p)-uniformly convex and having minimum in x_2 we have

$$\nabla f_2(x_1)^\top (x_1 - x_2) \geq f_2(x_1) - f_2(x_2) + \frac{\mu}{p}\|x_2 - x_1\|^p \geq \frac{\mu}{p}\|x_2 - x_1\|^p.$$

The first-order optimality of x_1 for f_1 implies that $\nabla f_1(x_1)^\top (x_1 - x_2) \leq 0$, thus

$$\begin{aligned}
\nabla f_2(x_1)^\top (x_1 - x_2) &= \nabla f_1(x_1)^\top (x_1 - x_2) + (\nabla f_2(x_1) - \nabla f_1(x_1))^\top (x_1 - x_2) \\
&\leq (\nabla f_2(x_1) - \nabla f_1(x_1))^\top (x_1 - x_2).
\end{aligned}$$

Applying the Hölder inequality to the above and combining together yields the result

$$\|x_2 - x_1\| \leq \left(\frac{p}{\mu}\|\nabla f_2(x_1) - \nabla f_1(x_1)\|_*\right)^{\frac{1}{p-1}}.$$

□

7.1 Proof of Lemma 3

Proof. Let $\Pi(y) = \arg\max_{x\in\mathcal{X}} \langle x, y\rangle - f(x)$. Take y_1, y_2 and $\gamma \in (0,1)$ and denote $x_1 = \Pi(y_1)$, $x_2 = \Pi(y_2)$, and $x_\gamma = \gamma x_1 + (1-\gamma)x_2$. From (Shalev-Shwartz, 2007, Lemma 15 (b)), we have $\gamma_1 \in \partial f(x_1)$ and $\gamma_2 \in \partial f(x_2)$. From f being (μ, p)-uniformly convex we get

$$f(x_\gamma) - f(x_1) - \langle y_1, x_\gamma - x_1\rangle \geq \frac{\mu}{p}\|x_\gamma - x_1\|^p$$
$$f(x_\gamma) - f(x_2) - \langle y_2, x_\gamma - x_2\rangle \geq \frac{\mu}{p}\|x_\gamma - x_2\|^p.$$

Adding γ times the first inequality to $(1-\gamma)$ times the second inequality yields

$$f(x_\gamma) - (\gamma f(x_1) + (1-\gamma)f(x_2)) + \gamma(1-\gamma)\langle y_2 - y_1, x_2 - x_1\rangle \geq \gamma(1-\gamma)\frac{\mu}{p}\|x_1 - x_2\|^p. \tag{8}$$

However, by f being (μ, p) uniformly convex we also have

$$f(x_\gamma) - (\gamma f(x_1) + (1-\gamma)f(x_2)) \leq -\gamma(1-\gamma)\frac{\mu}{p}\|x_1 - x_2\|^p. \tag{9}$$

Subtracting (9) from (8) yields

$$\langle y_2 - y_1, x_2 - x_1\rangle \geq \frac{2\mu}{p}\|x_2 - x_1\|^p,$$

which in combination with the Hölder inequality $\langle y_2 - y_1, x_2 - x_1\rangle \leq \|x_2 - x_1\|\|y_2 - y_1\|_*$ gives that

$$\frac{2\mu}{p}\|x_1 - x_2\|^{p-1} \leq \|y_1 - y_2\|_*$$

or

$$\|\nabla f^*(y_1) - \nabla f^*(y_2)\| \leq \left(\frac{p}{2\mu}\right)^{1/(p-1)} \|y_1 - y_2\|_*^{1/(p-1)}. \tag{10}$$

Let $T \in \mathbb{N}$. For all $t \in \{0, 1, \ldots, N\}$ define $\beta_t = t/T$. We have

$$\begin{aligned} f^*(y+\lambda) - f^*(y) &= f^*(y + \beta_T\lambda) - f^*(y + \beta_0\lambda) \\ &= \sum_{t=0}^{T-1} f^*(y + \beta_{t+1}\lambda) - f^*(y + \beta_T\lambda). \end{aligned}$$

From convexity of f^* we have

$$f^*(y + \beta_{t+1}) - f^*(y + \beta_t) \leq \langle \nabla f^*(y + \beta_{t+1}\lambda), (\beta_{t+1} - \beta_t)\lambda\rangle = \frac{1}{T}\langle \nabla f^*(y + \beta_{t+1}\lambda), \lambda\rangle$$

Thus we have

$$\begin{aligned} \langle \nabla f^*(y + \beta_{t+1}\lambda), \lambda\rangle &= \langle \nabla f^*(y), \lambda\rangle + \langle \nabla f^*(y + \beta_{t+1}) - \nabla f^*(y), \lambda\rangle \\ &\leq \langle \nabla f^*(y), \lambda\rangle + \|\nabla f^*(y + \beta_{t+1}\lambda) - \nabla f^*(y)\|\|\lambda\|_* \\ &\leq \langle \nabla f^*(y), \lambda\rangle + \left(\frac{p}{2\mu}\|\beta_{t+1}\lambda\|_*\right)^{1/(p-1)} \|\lambda\|_* \\ &= \langle \nabla f^*(y), \lambda\rangle + \left(\frac{p}{2\mu}\beta_{t+1}\right)^{1/(p-1)} \|\lambda\|_*^{p/(p-1)}, \end{aligned}$$

where the second inequality is the consequence of Hölders inequality and the third inequality comes from (10).

Now, we can express

$$f^*(y+\lambda) - f^*(y) \leq \langle \nabla f^*(y), \lambda\rangle + \left(\frac{p}{2\mu}\right)^{1/(p-1)} \|\lambda\|_*^{p/(p-1)} \frac{1}{T}\sum_{t=0}^{T-1} (\beta_{t+1})^{1/p-1}.$$

We are interested in the asymptotic bound as $T \to \infty$, for which we have that the limit is an Riemannian sum that can be expressed as an integral

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}(\beta_{t+1})^{1/p-1} = \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}\left(\frac{t}{T}\right)^{1/p-1} = \int_0^1 x^{\frac{1}{p-1}}\mathrm{d}x = \frac{p-1}{p}.$$

Thus we have

$$f^*(y+\lambda) - f^*(y) \le \langle \nabla f^*(y), \lambda\rangle + \frac{p-1}{p}\left(\frac{p}{2\mu}\right)^{1/(p-1)}\|\lambda\|_*^{p/(p-1)},$$

which after substituting $p = \frac{q}{q-1}$ and $\lambda = x - y$ yields the result. □

8 Proofs for results in Section 3

8.1 Proof of Lemma 5

Proof. Let $\hat{x}$ and $\tilde{x}$ be two $\hat{\varepsilon}$-minimizers of $f_S^{(\mu)}$ and $f_{S'}^{(\mu)}$ respectively that are (μ, p)-uniformly convex w.r.t. $\|\cdot\|$, where S and S' differ only in at most one sample. Let the minimum of $f_S^{(\mu)}$ be attained at $\hat{x}_\mu^*$ and the minimum of $f_{S'}^{(\mu)}$ is attained at $\tilde{x}_\mu^*$. By the triangle inequality we have

$$\|\hat{x} - \tilde{x}\| \le \|\hat{x}_\mu^* - \tilde{x}_\mu^*\| + \|\hat{x} - \hat{x}_\mu^*\| + \|\tilde{x} - \tilde{x}_\mu^*\|. \tag{11}$$

We denote the first term of the right hand side in (11) as S_1 and upper bound it

$$S_1 = \|\hat{x}_\mu^* - \tilde{x}_\mu^*\| \le \left(\frac{\nu}{n\mu}\|\nabla\ell(\hat{x}_\mu^*; z_i) - \nabla\ell(x_\mu^*; z_i')\|_*\right)^{1/(\nu-1)} \le \left(\frac{2\nu}{n\mu}G\right)^{1/(\nu-1)}, \tag{12}$$

where the first inequality follows from Lemma 17 applied to the regularized ERM functions that are uniformly convex w.r.t. $\|\cdot\|$ and the second inequality comes from the triangle inequality combined with the Lipschitz constant G for $\ell(\cdot; z)$ bounding the dual norm of $\nabla\ell(\hat{x}_\mu^*; z_i)$ and $\nabla\ell(\tilde{x}_\mu^*; z_i')$.

We denote the second term of the right hand side in (11) as S_2 and upper bound as follows

$$S_2 = \|\hat{x} - \hat{x}_\mu^*\| \le \left(\frac{\nu}{\mu}\left(f_S^{(\mu)}(\hat{x}) - f_S^{(\mu)}(\hat{x}_\mu^*)\right)\right)^{1/\nu} \le \left(\frac{\nu}{\mu}\hat{\varepsilon}\right)^{1/\nu}, \tag{13}$$

where the first inequality comes from the definition of $f_S^{(\mu)}(x)$ and the fact that x_μ^* is its minimizer, and the second inequality is the consequence of $\hat{x}$ being $\hat{\varepsilon}$-minimizer of $f_S^{(\mu)}$. Note that the third term of the right hand side in (11), which we denote as S_3, is bounded analogously as S_2 in (13).

When $\hat{\varepsilon} \le (\nu/\mu)^{1/(\nu-1)}(2G/n)^{\nu/(\nu-1)}$, we have that the upper bound on S_2 in (13) is smaller than the upper bound on S_1 in (12). In this case the stability is bounded for all $z \in \mathcal{Z}$ as

$$|\ell(\hat{x}; z) - \ell(\tilde{x}; z)| \le G\|\hat{x} - \tilde{x}\| \le 3G\left(\frac{2\nu}{n\mu}G\right)^{1/(\nu-1)}.$$

When the error of the regularized problem is small $\hat{\varepsilon} \le \mu D^\nu$, we can upper bound the optimization error of the non-regularized problem as

$$\begin{aligned} f_S(\hat{x}) - f_S(x^*) &= f_S^{(\mu)}(\hat{x}) - f_S^{(\mu)}(x^*) + \mu\psi(x^*) - \mu\psi(\hat{x}) \\ &\overset{①}{\le} f_S^{(\mu)}(\hat{x}) - f_S^{(\mu)}(x_\mu^*) + \mu\psi(x^*) - \mu\psi(x_0) \\ &\overset{②}{\le} \hat{\varepsilon} + \mu D^\nu \\ &\le 2\mu D^\nu, \end{aligned}$$

where ① holds by x_μ^* being the minimizer of $f_S^{(\mu)}$ while x_0 is the minimizer of $\psi(x)$ and ② comes from Lemma 16 and the definition of D. □

8.2 Proof of Theorem 6

Proof. Let $\mathcal{A}$ be an algorithm that for a dataset S returns an $\hat{\varepsilon}$-minimizer $\hat{x}$ of the regularized ERM $f_S^{(\mu)}$ with G-Lipschitz losses $\ell(\cdot, z)$. By the excess risk decomposition in (1) and Lemma 5 we have

$$\mathbb{E}_S\left[\delta f(\hat{x})\right] \leq \mathcal{E}_{\text{stab}}(\mathcal{A}, S) + \mathbb{E}_S\left[f_S(\hat{x}) - f_S(x^*)\right] \leq 3\left(\frac{2\nu}{n\mu}G^\nu\right)^{1/(\nu-1)} + 2\mu D^\nu,$$

where $x^* = \arg\min_{x\in\mathbb{R}^d} f_S(x)$.

To find the value of $\mu \geq 0$ that minimizes the upper bound we set its derivative in μ to zero, since the expression is convex on $\mu \geq 0$:

$$0 = 2D^\nu - 3\left(\frac{2\nu}{n}\right)^{\frac{1}{\nu-1}} G^{\frac{\nu}{\nu-1}} \frac{1}{\nu-1}\mu^{-\frac{\nu}{\nu-1}},$$

which after rearranging yields the minimum is attained for

$$\mu = \left(\frac{3}{\nu-1}\right)^{1-\frac{1}{\nu}} 2^{\frac{2}{\nu}-1}\nu^{\frac{1}{\nu-1}}\left(\frac{1}{n}\right)^{\frac{1}{\nu}} D^{1-\nu}G > 0 \tag{14}$$

and the value of the excess risk is upper bounded as

$$\begin{aligned}\mathbb{E}\left[\delta f(\hat{x})\right] \leq \mathcal{E}_{\text{stab}}(\mathcal{A}) + \mathcal{E}_{\text{opt}} &\leq 2^{1+\frac{2}{\nu}}\left(\frac{3}{\nu-1}\right)^{1-\frac{1}{\nu}}\left(\frac{1}{n}\right)^{\frac{1}{\nu}} DG \\ &\leq 8\left(\frac{1}{n}\right)^{\frac{1}{\nu}} DG,\end{aligned}$$

where in the first line we substituted μ from (14), and the second line follows from $2^{1+2/\nu}(\frac{3}{\nu-1})^{1-\frac{1}{\nu}} \leq 8$ for $\nu \geq 2$.

It remains to ensure the conditions of Lemma 5 are met, i.e. the error $\hat{\varepsilon}$ of the approximate solution $\hat{x}$, is small enough, and in particular bounded by the minimum of two terms. The bound on the first of the two terms in the minimum is met when

$$\begin{aligned}\hat{\varepsilon} \leq \mu D^\nu = 2^{2/\nu-1}\left(\frac{3}{\nu-1}\right)^{1-\frac{1}{\nu}}\left(\frac{1}{n}\right)^{\frac{1}{\nu}} DG \\ \leq 2\left(\frac{1}{n}\right)^{\frac{1}{\nu}} DG.\end{aligned}$$

where in the first line we substituted μ from (14), and the second line follows from $(\frac{3}{\nu-1})^{1-1/\nu} \leq 2$ and $2^{2/\nu-1} \leq 1$ for $\nu \geq 2$. The second error condition of Lemma 5 is

$$\begin{aligned}\hat{\varepsilon} \leq \left(\frac{\nu}{\mu}\right)^{1/(\nu-1)}\left(\frac{2G}{n}\right)^{\nu/(\nu-1)} &= 2^{2/\nu}\left(\frac{\nu-1}{3}\right)^{\frac{1}{\nu}} GD\left(\frac{1}{n}\right)^{1+\frac{1}{\nu}} \\ &\leq 6GD\left(\frac{1}{n}\right)^{1+\frac{1}{\nu}},\end{aligned}$$

where the second line follows from $2^{2/\nu}(\frac{\nu-1}{3})^{1/\nu} \leq 6$ for $\nu \geq 1$. □

8.3 Proof of Theorem 7

Proof. We consider two cases, when $1 \leq p \leq 1 + 1/\log(d)$ and when $p > 1 + 1/\log(d)$.

Case 1 ($1 \leq p \leq 1 + 1/\log(d)$)**:**
For $v \in \{\pm e_i\}_{i=1}^s$, where e_i's denote the first s canonical basis vectors, we define the distribution $Z \sim P_v$ as

$$Z_j = \begin{cases} 1, & \text{with probability } \frac{1+\delta v_j}{2} \\ -1, & \text{with probability } \frac{1-\delta v_j}{2} \end{cases} \quad \text{for } j \in \{1, \ldots, sbeginproof$$

and $Z_j = 0$ for $j \in \{s+1, \dots, d\}$. The distribution is supported only on the first $s \leq d$ entries and is well defined when $\delta \leq 1$. We have that $\mathbb{E}_{z\sim P_v} = \delta v$.

Let $\ell(x; z) = Gs^{-1/p^*} x^\top z$ be a linear loss where $1/p + 1/p^* = 1$. The population loss for $v \sim P_v$ is

$$\ell_v(x) = \mathbb{E}_{z\sim P_v} \ell(x; z) = \delta s^{-1/p^*} x^\top z$$

and its infimum for $x \in \mathcal{B}_{\|\cdot\|_p}(R)(r)$ is

$$\ell_v^* := \inf_{x\in\mathcal{B}_{\|\cdot\|_p}(R)(r)} \ell_v(x) = -\delta s^{-1/p^*}.$$

For $v \neq v'$ we have

$$\begin{aligned}
\mathrm{d}_{\mathrm{opt}}(v, v', \mathcal{B}_{\|\cdot\|_p}(r)) = \inf_{x\in\mathcal{B}_{\|\cdot\|_p}(r)} \ell_v(x) + \ell_{v'}(x) - \ell_v^* - \ell_{v'}^* &= s^{-1/p}\delta \inf\left((v+v')^\top x + 2\right) \\
&= \delta s^{-1/p^*}(2 - \|v + v'\|_{p^*}) \\
&= (2 - \sqrt{2})\delta s^{-1/p^*}.
\end{aligned}$$

From (Levy and Duchi, 2019, Lemma 2, Sec. A.1) combined with Fano's inequality given in (Levy and Duchi, 2019, Proposition 7) yields a lower bound on the excess risk as

$$\inf_{\hat{x}\in\mathcal{B}_{\|\cdot\|_p}(r)} \sup_{P\in\mathcal{P}} \mathbb{E}_{S\sim P^n}\left[\delta f(\hat{x})\right] \geq \frac{r}{4} G\delta s^{-1/p^*} \left(1 - \frac{3n\delta^2 + \log 2}{\log(2s)}\right). \tag{15}$$

Set $\delta = \sqrt{\log(s)/(6n)}$, which yields

$$\begin{aligned}
\inf_{\hat{x}\in\mathcal{B}_{\|\cdot\|_p}(r)} \sup_{P\in\mathcal{P}} \mathbb{E}_{S\sim P^n}\left[\delta f(\hat{x})\right] &\geq \frac{r}{4} G s^{-1/p^*} \sqrt{\frac{\log(s)}{n}} \left(1 - \frac{\frac{32}{6}\log(s) + \frac{4}{3}\log 2}{s}\right) \\
&\geq \frac{r}{4} G s^{-1/p^*} \sqrt{\frac{\log(s)}{n}},
\end{aligned}$$

where the second inequality holds for $s \geq 8$ which ensures that the right hand side remains positive. Note that for $p \leq 1 + 1/\log(d)$ and $s \leq d$, we have that $p \leq 1 + 1/\log(s)$, for which we can bound $s^{-1/p^*} = s^{-(p-1)/p} \geq 1/e$ completing the proof

$$\inf_{\hat{x}\in\mathcal{B}_{\|\cdot\|_p}(r)} \sup_{P\in\mathcal{P}} \mathbb{E}_{S\sim P^n}\left[\delta f(\hat{x})\right] \geq \frac{r}{4e} G\sqrt{\frac{\log(s)}{n}} \geq \frac{rG}{12}\sqrt{\frac{\log(s)}{n}}$$

when we choose $s = d$. In order for $\delta \leq 1$, we need that $\log(d) \leq 6n$. When $\log(d) \geq 6n$, we choose $\delta = 1$ and get

$$\inf_{\hat{x}\in\mathcal{B}_{\|\cdot\|_p}(r)} \sup_{P\in\mathcal{P}} \mathbb{E}_{S\sim P^n}\left[\delta f(\hat{x})\right] \geq \frac{rG}{12}.$$

Case 2 ($p > 1 + 1/\log(d)$)**:**
Let $\ell(x; z) = Gx^\top z$ be a linear loss. For $v \in \{\pm 1\}^d$ we define the distribution $Z \sim P_v$ as

$$Z = \begin{cases} v_j e_j, & \text{with probability } \frac{1+\delta}{2s} \\ -v_j e_j, & \text{with probability } \frac{1-\delta}{2s} \end{cases} \quad \text{for } j \in \{1, \dots, s\},$$

supported only on the first $s \leq d$ entries. The distribution is well defined when $\delta \leq 1$.

We have that

$$\begin{aligned}
\ell_v(x) = \mathbb{E}_{z\sim P_v}\left[\ell(x; z)\right] &= \sum_{j=1}^{s} \frac{1+\delta}{2s}\ell(x, v_j e_j) + \frac{1-\delta}{2s}\ell(x, -v_j e_j) \\
&= G\sum_{j=1}^{s} \frac{1}{2s} x_j v_j \left(1 + \delta - 1 + \delta\right) \\
&= \frac{G\delta}{s} x^\top v_\Omega,
\end{aligned}$$

where we denote v_Ω to be a vector that contains entries of v at indices $\Omega = \{1, \dots s\}$ and zeroes otherwise. By duality the minimum of this function can be computed as

$$\ell_v^* = \min_{x \in \mathcal{B}_{\|\cdot\|_p}(r)} \frac{G\delta}{d} x^\top v_\Omega = \min_{x \in \mathcal{B}_{\|\cdot\|_p}(1)} \frac{G\delta r}{d} x^\top v_\Omega = -\frac{G\delta r}{s} \|v_\Omega\|_{p^*},$$

where $1/p^* + 1/p = 1$.

For $v, v' \in \{\pm 1\}^d$, we have

$$\begin{aligned}
\mathrm{d}_{\mathrm{opt}}(v, v', \mathcal{B}_{\|\cdot\|_p}(r)) :&= \inf_{x \in \mathcal{B}_{\|\cdot\|_p}(r)} \ell_v(x) - \ell_v^* + \ell_{v'}(x) - \ell_{v'}^* \\
&= \inf_{x \in \mathcal{B}_{\|\cdot\|_p}(1)} G\frac{\delta r}{s} \left(x^\top (v_\Omega + v'_\Omega) + \|v_\Omega\|_{p^*} + \|v'_\Omega\|_{p^*} \right) \\
&= G\frac{\delta r}{s} \left(\|v_\Omega\|_{p^*} + \|v'_\Omega\|_{p^*} - \|v_\Omega + v'_\Omega\|_{p^*} \right) \\
&= 2G\frac{\delta r}{s} \left(s^{1/p^*} - (s - \|v_\Omega - v'_\Omega\|_0)^{1/p^*} \right),
\end{aligned}$$

where $\|v_\Omega - v'_\Omega\|_0$ denotes the ℓ_0 norm counting the number of different entries between v_Ω and v'_Ω i.e., the Hamming distance between v_Ω and v'_Ω.

Now, it is sufficient to provide packing of $\left\{v \in \{\pm 1\}^d : \mathrm{supp}(v) \subseteq \Omega\right\}$ that restricts $\|v_\Omega - v'_\Omega\|_0$ since, if v, v' are not supported on Ω we have that $\mathrm{d}_{\mathrm{opt}}(v, v', \mathcal{B}_p) = 0$.

It is sufficient to restrict $\|v_\Omega - v'_\Omega\|_1 \geq \frac{s}{2}$ as this implies also that $\|v_\Omega - v'_\Omega\|_0 \geq \frac{s}{2}$. We can use the Gilbert-Varshimov bound (Duchi, 2023, Lemma 7.5), that gives an $\frac{s}{2}$ ℓ_1-packing of $\left\{v \in \{\pm 1\}^d : \mathrm{supp}(v) \subseteq \Omega\right\}$ of size at least $\exp(\mathrm{s}/8)$. Let $\mathcal{V}$ be the packing, we have that

$$\forall v, v' \in \mathcal{V} \quad \text{s.t.} \quad v \neq v' : \mathrm{d}_{\mathrm{opt}}(v, v', \mathcal{B}_{\|\cdot\|_p}(r)) \geq \frac{r}{2} G\delta s^{-1/p},$$

where we used that $s^{1/p^* - 1} = s^{-1/p}$.

By (Levy and Duchi, 2019, Lemma 2, Sec. A.1), we have

$$\inf_{\hat{x} \in \mathcal{B}_{\|\cdot\|_p}(r)} \sup_{P \in \mathcal{P}} \mathbb{E}_{S \sim P^n} \left[\delta f(\hat{x})\right] \geq \frac{r}{4} G\delta s^{-1/p} \left(1 - \frac{3n\delta^2 + \log 2}{s/8}\right). \tag{16}$$

Set $\delta = \sqrt{\frac{s}{48n}}$. Since $s \leq 48n$ we have

$$\inf_{\hat{x} \in \mathcal{B}_{\|\cdot\|_p}(r)} \sup_{P \in \mathcal{P}} \mathbb{E}_{S \sim P^n} \left[\delta f(\hat{x})\right] \geq \frac{r}{16} G \frac{s^{1/2 - 1/p}}{\sqrt{n}},$$

when $s \geq 32 \log(2)$. In the low-dimensional case, when $d \leq 48n$, we can set $s = d$ to recover the lower bound in (Levy and Duchi, 2019).

In the high-dimensional case, when $d > 48n$ and $1 + \log^{-1}(d) \leq p \leq 2$ we choose $s = 1$ to get a lower bound as

$$\inf_{\hat{x} \in \mathcal{B}_{\|\cdot\|_p}(r)} \sup_{P \in \mathcal{P}} \mathbb{E}_{S \sim P^n} \left[\delta f(\hat{x})\right] \geq \frac{r}{16} \frac{G}{\sqrt{n}}.$$

Otherwise, when $d > 48n$ and $p \geq 2$, we can set $s = n$, for which $\delta = \sqrt{\frac{1}{48}} \leq 1$, to get that

$$\inf_{\hat{x} \in \mathcal{B}_{\|\cdot\|_p}(r)} \sup_{P \in \mathcal{P}} \mathbb{E}_{S \sim P^n} \left[\delta f(\hat{x})\right] \geq \frac{r}{16} G n^{-1/p}.$$

□

9 Proofs for results in Section 4

9.1 Proof of Theorem 8

Proof. Let $f_S^{(\mu)}(x) = f_S(x) + \mu\frac{\alpha}{p}\|x - x_0\|_p^p$ be the regularized ERM. From the theorem's assumption we have that $\|x^* - x_0\|_p \leq R$ for $x^* = \arg\min_{x\in\mathcal{X}} f_S(x)$, which implies by Lemma 16 also that $\|x_\mu^* - x_0\|_p \leq R$ for $x_\mu^* = \arg\min_{x\in\mathbb{R}^d} f_S^{(\mu)}(x)$. Based on the value of p, we consider two cases:

Case 1 ($p \geq 2$): When $p \geq 2$, by the definition of $\alpha = 2^{p-2}$ the regularization term $\mu\frac{\alpha}{p}\|x - x_0\|_p^p$ is (μ, p)-uniformly convex and, by Lemma 14, it is also $(p-1)\mu\alpha R^{p-2}$-smooth in $\mathcal{B}_{\|\cdot\|_p}(x_0, R)$ w.r.t ℓ_p-norm. As a result $f_S^{(\mu)}(x) = f_S(x) + \mu\frac{\alpha}{p}\|x - x_0\|_p^p$ is (μ, p)-uniformly convex and $\hat{L}$-smooth where $\hat{L} = L + (p-1)\mu\alpha R^{p-2}$. Assume $\mu \leq LR^{2-p}/((p-1)\alpha)$ that implies $\hat{L} \leq 2L$ and which we later show is satisfied for our choice of μ when T is large enough.

We start with x_0 for which we know that $\|x_0 - x^*\|_p \leq R_0$, and by Lemma 16, also that $\|x_0 - x_\mu^*\|_p \leq R_0$, where $R_0 \overset{\text{def}}{=} R$. From $f_S^{(\mu)}$ being $2L$-smooth we have $f_S^{(\mu)}(x_0) - f_S^{(\mu)}(x_\mu^*) \leq \hat{\varepsilon}_0 \overset{\text{def}}{=} LR_0^2$.

The algorithm will take $i = 1, \ldots, r$ stages to achieve the final accuracy $\hat{\varepsilon}$. At the stage i, the algorithm starts with an estimate x_{i-1} within the distance $\|x_{i-1} - x_\mu^*\|_p \leq R_{i-1}$ and will output a point x_i that achieves accuracy $\hat{\varepsilon}_i$, such that $\hat{\varepsilon}_i = \hat{\varepsilon}_{i-1}/2$. From the (μ, p)-uniform convexity of $f_S^{(\mu)}$ combined with the guaranteed accuracy of x_i on $f_S^{(\mu)}$, we then have that the output x_i of stage i satisfies

$$\hat{\varepsilon}_i \geq f_S^{(\mu)}(x_i) - f_S^{(\mu)}(x_\mu^*) \geq \frac{\mu}{p}\|x_i - x_\mu^*\|_p^p.$$

Thus, from the uniform convexity of $f_S^{(\mu)}(x)$, at the next stage, $i+1$, the algorithm is initialized with a point x_i for which $\|x_i - x_\mu^*\|_p \leq R_i \overset{\text{def}}{=} (p\hat{\varepsilon}_i/\mu)^{\frac{1}{p}}$. We can initialize the algorithm at the next stage with this R_i.

To achieve $\hat{\varepsilon}$ accuracy, it is sufficient to run the algorithm for $r = \log_2(\hat{L}R_0^2/(2\hat{\varepsilon})) = \log_2(LR_0^2/\hat{\varepsilon})$ stages, which amounts to the following number of gradient oracle calls

$$\begin{aligned}
T = \sum_{i=1}^{r} \hat{T}_i &= C\sum_{i=1}^{r}\left(\frac{\hat{L}R_{i-1}^2}{\hat{\varepsilon}_i}\right)^{\frac{1}{\gamma}} \\
&= C\sum_{i=0}^{r-1}\left(\hat{L}\left(\frac{p}{\mu}\right)^{\frac{2}{p}}\left(\frac{1}{\hat{\varepsilon}_i}\right)^{1-\frac{2}{p}}\right)^{\frac{1}{\gamma}} \\
&\leq C\hat{L}^{\frac{1}{\gamma}}\left(\frac{p}{\mu}\right)^{\frac{2}{p\gamma}}\sum_{i=0}^{r-1}\left(\frac{1}{2^{r-i-1}\hat{\varepsilon}}\right)^{\left(1-\frac{2}{p}\right)\frac{1}{\gamma}} \\
&= C\hat{L}^{\frac{1}{\gamma}}\left(\frac{p}{\mu}\right)^{\frac{2}{p\gamma}}\left(\frac{1}{\hat{\varepsilon}}\right)^{\left(1-\frac{2}{p}\right)/\gamma}\sum_{i=0}^{r-1}\left(\frac{1}{2^{r-i-1}}\right)^{\left(1-\frac{2}{p}\right)\frac{1}{\gamma}},
\end{aligned}$$

where for the inequality we used that $\hat{\varepsilon}_i = \hat{\varepsilon}_{r-1}2^{r-1-i} \geq \hat{\varepsilon}2^{r-i-1}$ for all $i \in \{1, \ldots, r\}$.

It remains to evaluate the sum

$$\sum_{i=0}^{r-1}\left(\left(\frac{1}{2}\right)^{\left(1-\frac{2}{p}\right)/\gamma}\right)^{r-i-1} \leq \sum_{i=0}^{r-1}\left(1^{1/\gamma}\right)^{r-i-1} = r = \log_2(LR^2/\hat{\varepsilon})$$

since for a fixed $\gamma > 0$ we have $(\frac{1}{2})^{(1-2/p)/\gamma} \leq 1$ when $p \geq 2$.

As a result, to reach accuracy $\hat{\varepsilon}$, it is sufficient to make the following number of gradient oracle calls

$$T = C\log_2\left(\frac{LR^2}{\hat{\varepsilon}}\right)p^{\frac{2}{p\gamma}}L^{\frac{1}{\gamma}}\mu^{-\frac{2}{p\gamma}}\hat{\varepsilon}^{-\left(1-\frac{2}{p}\right)/\gamma}, \tag{17}$$

where we used that $\hat{L} \leq 2L$ when $\mu \leq LR^{2-p}/((p-1)\alpha)$.

From Lemma 5 with $G = 2LR$ we have that, in order for the output to be uniformly stable, we need to achieve accuracy $\hat{\varepsilon}$ satisfying both $\hat{\varepsilon} \leq \mu D^p = \mu \frac{\alpha}{p} R^p$ and $\hat{\varepsilon} \leq (p/\mu)^{1/(p-1)} \left(\frac{4LR}{n}\right)^{p/(p-1)}$.

When $\mu \leq \alpha^{\frac{1-p}{p}} R^{2-p} \frac{4pL}{n}$, we have that $\mu \frac{\alpha}{p} R^p \leq (p/\mu)^{1/(p-1)} \left(\frac{4LR}{n}\right)^{p/(p-1)}$ and thus the accuracy we require is $\hat{\varepsilon} \leq \mu \frac{\alpha}{p} R^p$. To reach an accuracy that is upper bounded as $\hat{\varepsilon} \leq \mu \frac{\alpha}{p} R^p$ we get from (17) that the number of iterations must be at least

$$T \geq C \log_2 \left(\frac{LR^{2-p} p}{\mu \alpha} \right) p^{1/\gamma} L^{\frac{1}{\gamma}} \mu^{-\frac{1}{\gamma}} \alpha^{\frac{2-p}{p\gamma}} R^{\frac{2-p}{\gamma}}. \tag{18}$$

To ensure (18) holds for a given T, we choose

$$\mu = \left(\frac{C}{T}\right)^{\gamma} pL\alpha^{\frac{2-p}{p}} R^{2-p} \log_2 \left(\left(\frac{T}{C}\right)^{\gamma} \alpha^{-2/p} \right)^{\gamma},$$

which satisfies (18) when $T \geq C(2\alpha^{2/p})^{1/\gamma}$. For this choice of μ, the requirement that $\mu \leq \alpha^{\frac{1-p}{p}} R^{2-p} \frac{4pL}{n}$ is met when $T = \tilde{\Omega}_p \left(n^{1/\gamma}\right)$. To ensure that $\hat{L} \leq 2L$, we need $\mu \leq LR^{2-p}/((p-1)\alpha)$, which for our choice of μ is satisfied when $T \geq C \left(p(p-1)\alpha^{\frac{2}{p}}\right)^{1/\gamma}$. Joining the two conditions gives

$$T = \tilde{\Omega}_p \left(n^{1/\gamma}\right).$$

For this μ the accuracy $\hat{\varepsilon}$ we wish to reach from Lemma 5 is

$$\hat{\varepsilon} \leq \mu \frac{\alpha}{p} R^p = \alpha^{2/p} LR^2 \left(\frac{C}{T}\right)^{\gamma} \log_2 \left(\left(\frac{T}{C}\right)^{\gamma} \alpha^{-2/p} \right)^{\gamma},$$

which gives us the requirement for the minimum number of stages $r = \log_2(LR^2/\hat{\varepsilon})$.

By Lemma 5 with $G = 2LR$, for this range of T, the optimization error will be

$$\mathcal{E}_{\text{opt}} = f_S(x_T) - f_S(x^*) \leq \mu \frac{\alpha}{p} R^p = \left(4\frac{C}{T}\right)^{\gamma} \log_2 \left(\left(\frac{T}{C}\right)^{\gamma} \alpha^{-2/p} \right)^{\gamma} \alpha^{2/p} LR^2.$$

and the stability is

$$\begin{aligned} \mathcal{E}_{\text{stab}}(\mathcal{A}) \leq 3 \left(\frac{2p}{n\mu} G^p\right)^{\frac{1}{p-1}} &= 3 \left(2L^{p-1} p^{p-1} \alpha^{\frac{p-2}{p}} R^{2(p-1)} \left(\frac{T}{4C}\right)^{\gamma} \log_2 \left(\left(\frac{T}{C}\right)^{\gamma} \alpha^{-2/p} \right)^{-\gamma} \frac{1}{n} \right)^{\frac{1}{p-1}} \\ &= \tilde{\mathcal{O}}_p \left(\left(\frac{T^{\gamma}}{n}\right)^{\frac{1}{p-1}} LR^2 \right), \end{aligned}$$

when $T = \tilde{\Omega}_p \left(n^{1/\gamma}\right)$.

Case 2 ($p \in (1,2)$): When $p \in (1,2)$, by the choice of $\alpha = 2^{p-3}(1-1/p)^{p-1}$ and Lemma 15, we have that $\mu \frac{\alpha}{p} \|x\|_p^p$ is (μ, p)-Hölder smooth globally w.r.t ℓ_p-norm. From Lemma 14 we have that $\mu \frac{\alpha}{p} \|x\|_p^p$ is $\mu\alpha R^{p-2}(p-1)$-strongly convex inside of $x \in \mathcal{B}_{\|\cdot\|_p}(R)$ w.r.t ℓ_p-norm.

In the following, denote $\psi(x) = \mu \frac{\alpha}{p} \|x - x_0\|_p^p$ for ease of notation. We derive the combined Hölder smoothness of the regularized ERM as follows

$$\begin{aligned} f_S(x) + \psi(x) - (f_S(y) + \psi(y) + \langle \nabla f_S(y) + \nabla \psi(x), x - y \rangle) &\leq \frac{L}{2} \|x - y\|_p^2 + \frac{\mu}{p} \|x - y\|_p^p \\ &\leq \frac{1}{p} \left(p 2^{1-p} LR^{2-p} + \mu \right) \|x - y\|_p^p, \end{aligned}$$

for all $x, y \in \mathcal{B}_{\|\cdot\|_p}(R)$. Consequently, we have that $f_S^{(\mu)}(x)$ is $(\hat{L}, p)$-Hölder smooth globally, where $\hat{L} = p2^{1-p} LR^{2-p} + \mu$ and $\mu\alpha R^{p-2}(p-1)$-strongly convex for $x \in \mathcal{B}_{\|\cdot\|_p}(R)$ w.r.t ℓ_p-norm. Assume $\mu \leq p2^{1-p} LR^{2-p}$ that implies $\hat{L} \leq p2^{2-p} LR^{2-p}$ and which we later show is satisfied for our choice of μ when T is large enough.

We start with x_0 for which we know that $\|x_0 - x_\mu^*\|_p \le R_0 \stackrel{\text{def}}{=} R$. By Hölder smoothness of $f_S^{(\mu)}$ we have that $f_S^{(\mu)}(x_0) - f_S^{(\mu)}(x_\mu^*) \le \hat{\varepsilon}_0 \stackrel{\text{def}}{=} \hat{L}R_0^p/p$.

The algorithm will take $i = 1, \ldots, r$ stages to achieve final accuracy $\hat{\varepsilon}$. At the stage i, the algorithm starts with an estimate x_{i-1} within the distance $\|x_{i-1} - x_\mu^*\|_p \le R_{i-1}$ and will output a point x_i that achieves accuracy $\hat{\varepsilon}_i$, such that $\hat{\varepsilon}_i = \hat{\varepsilon}_{i-1}/2$. From $\mu\alpha R^{p-2}(p-1)$-strong convexity of $f_S^{(\mu)}$ we have that the output x_i of stage i satisfies

$$\hat{\varepsilon}_i \ge f_S^{(\mu)}(x_i) - f_S^{(\mu)}(x_\mu^*) \ge \mu\alpha R^{p-2}(p-1)\|x_i - x_\mu^*\|_p^2.$$

Thus, from the strong convexity of $f_S^{(\mu)}$, at the next stage, $i+1$, the algorithm is initialized with a point x_i for which $\|x_i - x_\mu^*\|_p \le R_i \le (R^{2-p}\hat{\varepsilon}_i\mu^{-1}\alpha^{-1}(p-1)^{-1})^{1/2}$.

To achieve $\hat{\varepsilon}$ accuracy, it is sufficient to run the algorithm for $r = \log_2(\hat{L}R_0^p/(p\hat{\varepsilon}))$ stages, which amounts to the following number of gradient oracle calls

$$\begin{aligned}
T = \sum_{i=1}^{r} \hat{T}_i &= C\sum_{i=0}^{r-1}\left(\frac{\hat{L}R_i^p}{\hat{\varepsilon}_i}\right)^{\frac{1}{\gamma}} \\
&= C\sum_{i=0}^{r-1}\left(\hat{L}\left(\frac{R^{2-p}}{\mu\alpha(p-1)}\right)^{\frac{p}{2}}\left(\frac{1}{\hat{\varepsilon}_i}\right)^{1-\frac{p}{2}}\right)^{\frac{1}{\gamma}} \\
&\le C\hat{L}^{\frac{1}{\gamma}}\left(\frac{R^{2-p}}{\mu\alpha(p-1)}\right)^{\frac{p}{2\gamma}}\sum_{i=0}^{r-1}\left(\frac{1}{2^{r-i-1}\hat{\varepsilon}}\right)^{(1-\frac{p}{2})\frac{1}{\gamma}} \\
&= C\hat{L}^{\frac{1}{\gamma}}\left(\frac{R^{2-p}}{\alpha\mu(p-1)}\right)^{\frac{p}{2\gamma}}\left(\frac{1}{\hat{\varepsilon}}\right)^{(1-\frac{p}{2})/\gamma}\sum_{i=0}^{r-1}\left(\frac{1}{2^{r-i-1}}\right)^{(1-\frac{p}{2})\frac{1}{\gamma}}
\end{aligned}$$

where for the inequality we used that $\hat{\varepsilon}_i = \hat{\varepsilon}_{r-1}2^{r-1-i} \ge \hat{\varepsilon}2^{r-i-1}$ for all $i \in \{1, \ldots, r\}$.

It remains to evaluate the sum

$$\sum_{i=0}^{r-1}\left(\left(\frac{1}{2}\right)^{(1-\frac{p}{2})/\gamma}\right)^{r-i-1} \le \sum_{i=0}^{r-1}\left(1^{1/\gamma}\right)^{r-i-1} = r = \log_2(LR^p/(p\hat{\varepsilon}))$$

since for a fixed $\gamma > 0$ we have $(\frac{1}{2})^{(1-p/2)/\gamma} \le 1$ when $p \in (1, 2)$.

As a result, to reach accuracy $\hat{\varepsilon}$, it is sufficient to make the following number of gradient oracle calls

$$T = \log_2\left(\frac{LR^p}{p\hat{\varepsilon}}\right) C\left(\frac{2p}{(p-1)^{p/2}}\right)^{1/\gamma} L^{1/\gamma}R^{\frac{(2-p)(2+p)}{2\gamma}}\alpha^{-\frac{p}{2\gamma}}\mu^{-\frac{p}{2\gamma}}\hat{\varepsilon}^{-(1-\frac{p}{2})/\gamma}, \tag{19}$$

where we used that $\hat{L} \le p2^{2-p}R^{2-p}L$ and that $2^{1-p} \le 2$ for $p \in (1, 2)$

From Lemma 5 with $G = 2LR$, we have that for a $\mu\alpha R^{p-2}(p-1)$-strongly convex regularizer to result in an output that is uniformly stable, we need to achieve accuracy $\hat{\varepsilon}$ satisfying both $\hat{\varepsilon} \le \mu\alpha R^{p-2}(p-1)D^2 = \mu\alpha^{1+2/p}p^{-2/p}(p-1)R^p$ and $\hat{\varepsilon} \le 32L^2R^{4-p}\mu^{-1}\alpha^{-1}(p-1)^{-1}n^{-2}$.

When $\mu \le 2^{2+\frac{1}{2}}LR^{2-p}(p^{1/p}/(p-1))\alpha^{-\frac{p+1}{p}}/n$, we have that $\mu\alpha^{1+2/p}p^{-2/p}(p-1)R^p \le 32L^2R^{4-p}\mu^{-1}\alpha^{-1}(p-1)^{-1}n^{-2}$ and thus the accuracy we need is $\hat{\varepsilon} \le \mu\alpha^{1+2/p}p^{-2/p}(p-1)R^p$. To reach an accuracy that is upper bounded as $\hat{\varepsilon} \le \mu\alpha^{1+2/p}p^{-2/p}(p-1)R^p$ we get from (19) that it is sufficient to perform a number of iterations at least

$$T \ge C2^{1/\gamma}\left(\frac{p^{2/p}}{p-1}\right)^{\frac{1}{\gamma}}R^{\frac{2-p}{\gamma}}L^{\frac{1}{\gamma}}\mu^{-\frac{1}{\gamma}}\alpha^{-\frac{2}{p\gamma}}\log_2\left(\frac{Lp^{2/p-1}}{(p-1)\mu\alpha^{(1+2/p)}}\right). \tag{20}$$

To ensure (20) holds for a given T, we choose

$$\mu = \left(\frac{C}{T}\right)^{\gamma}\frac{4}{p-1}LR^{2-p}\alpha^{-\frac{2}{p}}\log_2\left(\frac{1}{2p\alpha}\left(\frac{T}{C}\right)^{\gamma}\right)^{\gamma}, \tag{21}$$

where we used that $p^{p/2} \leq 2$ for $p \in (1,2)$ and require $T \geq C(4p\alpha)^{1/\gamma}$.

The choice of μ in (21) satisfies the requirement $\mu \leq 2^{2+\frac{1}{2}} LR^{2-p}(p^{1/p}/(p-1))\alpha^{-\frac{p+1}{p}}/n$ when $T = \tilde{\Omega}_p(n^{1/\gamma})$. In order to satisfy $\mu \leq p2^{1-p}LR^{2-p}$ which we need to have $\hat{L} \leq p2^{2-p}LR^{2-p}$, for the choice of μ in (21), we need $T \geq 2^{3(1+1/\gamma)}C^{1/\gamma}\alpha^{-\frac{2}{p\gamma}}/(p-1)^{1/\gamma}$.

With the choice of μ as in (21), we aim to achieve accuracy

$$\hat{\varepsilon} \leq \mu\alpha^{1+2/p}p^{-2/p}(p-1)R^p = 2\left(\frac{C}{T}\right)^{\gamma}\log_2\left(\frac{1}{2p\alpha}\left(\frac{T}{C}\right)^{\gamma}\right)^{\gamma}\alpha LR^2,$$

which gives us the requirement for the minimum number of stages $r = \log_2(LR^p/(p\hat{\varepsilon})) = \log_2(R^{p-2}/(2p\alpha))$. From Lemma 5, we have that the optimization error will be

$$\mathcal{E}_{\text{opt}} = f_S(x_T) - f_S(x^*) \leq 4\left(\frac{C}{T}\right)^{\gamma}\alpha LR^2\log_2\left(\frac{1}{2p\alpha}\left(\frac{T}{C}\right)^{\gamma}\right)^{\gamma}$$

and the stability is

$$\mathcal{E}_{\text{stab}}(\mathcal{A}) = \tilde{\mathcal{O}}_p\left(\frac{T^{\gamma}LR^2}{n}\right),$$

when $T = \tilde{\Omega}_p(n^{1/\gamma})$. □

9.2 Proof of Corollary 10

Proof. Let $p > 1$ and denote $\hat{p} = \max\{p, 2\}$. Let $\psi(x) = \frac{1}{2}\|x - x_0\|_2^2$. When $x^* \in \mathcal{B}_{\|\cdot\|_p}(x_0, R)$, we have due to the choice of $\psi(x)$ that

$$\|x^* - x_0\|_p \geq \sqrt{2}d^{1/\hat{p}-1/2}\left(\frac{1}{2}\|x^* - x_0\|_2^2\right)^{1/2},$$

which implies that we can apply Theorem 6 with $D = \sqrt{1/2}d^{1/2-1/\hat{p}}R$.

The regularized objective $f_S^{(\mu)}$ is $(L+\mu)$-smooth. Assume that $\mu \leq L$, which we soon show that it holds for n large enough, such that the objective $f_S^{(\mu)}$ is $2L$-smooth.

We apply Theorem 6 using the fact that $G = 2LR$ for $x \in \mathcal{B}_{\|\cdot\|_p}(x_0, R)$ as explained in (4), which requires to choose $\mu = 2^{1-1/2}\sqrt{3}D^{-1}G = 2\sqrt{3}d^{1/\hat{p}-1/2}n^{-1/2}L$ where G-Lipschitz and L-smooth constants of the loss are w.r.t ℓ_p-norm. To ensure that $\mu \leq L$ we assumed earlier, we need $n \geq 12d^{2/\hat{p}-1}$.

By the developments in the convergence analysis of USOLP in Section 9.1 for $p = 2$, the equation (17) gives that the total number of gradient oracle calls for this choice of μ is

$$T = \tilde{\Omega}_p\left(\left(\frac{n^{1/2}}{d^{1/\hat{p}-1/2}}\right)^{1/\gamma}\right).$$

and the bound on the excess risk is

$$\mathbb{E}_S\left[\delta f(\hat{x}_\mu)\right] = \mathcal{O}_p\left(LR^2\frac{d^{1/\hat{p}-1/2}}{n^{1/2}}\right),$$

where we used that $D = \sqrt{1/2}d^{1/2-1/\hat{p}}R$. □

9.3 Proof of Corollary 11

Proof. **Case 1** ($p \geq 2$)**:** We choose $\mu = (\frac{3}{p-1})^{1-\frac{1}{p}}2^{2/p}p^{1+\frac{1}{p(p-1)}}\alpha^{\frac{1-p}{p}}\left(\frac{1}{n}\right)^{\frac{1}{p}}R^{2-p}L$ as given in the proof in (14) with $G = 2LD$ and $D = (\alpha/p)^{1/p}R$. When $n \geq 3^{p-1}4(p-1)p^{p+\frac{1}{p-1}}\alpha$, we have that $\mu \leq LR^{2-p}/((p-1)\alpha)$ and thus $\hat{L} \leq 2L$ as given in the proof of Theorem 8 for $p \geq 2$, where $\hat{L}$ is the smoothness of the regularized empirical risk objective.

By the developments in the convergence analysis of USOLP given in Section 9.1 for case $p \geq 2$, the equation (17) gives that the number of iterations for x_T to be $\hat{\varepsilon}$-accurate on $f_S^{(\mu)}$ is

$$T = \tilde{\Omega}_p \left(L^{\frac{1}{\gamma}(1-\frac{2}{p})} R^{2\frac{p-2}{p\gamma}} \alpha^{2\frac{p-1}{p^2\gamma}} n^{\frac{2}{p^2\gamma}} \hat{\varepsilon}^{-(1-\frac{2}{p})/\gamma} \right),$$

which, since we need to ensure $\hat{\varepsilon} \leq 6L(\alpha/p)^{1/p}R^2/n^{1+1/p}$ in order for Theorem 6 to apply, implies we need to have number of iterations as

$$T = \tilde{\Omega}_p \left(\left(n^{1-1/p} \right)^{1/\gamma} \right).$$

Consequently, Theorem 6 upper bounds the excess risk of x_T as $\mathbb{E}_S[\delta f(x_T)] \leq 16(\alpha/p)^{1/p} L R^2/n^{1/p}$ for $p \geq 2$.

Case 2 ($p \in (1,2)$): We have that $f_S^{(\mu)}$ is $(\hat{L}, p)$-Hölder smooth and $\mu\alpha(p-1)R^{p-2}$-strongly convex in $\mathcal{B}_{\|\cdot\|_p}(x_0, R)$. We set $\mu = 4\sqrt{3}\alpha^{-3/2}\sqrt{\frac{1}{n}}R^{2-p}L/(p-1)$, which is equivalent to $\mu(p-1)D^{p-2} = 4\sqrt{3}\sqrt{\frac{1}{n}}L$, and corresponds to the optimal choice of the strong-convexity constant in Theorem 6 when the regularizer is $\mu\alpha(p-1)R^{p-2}$-strongly convex. By the developments in the convergence analysis of USOLP given in Section 9.1 for case $p \in (1,2)$, the equation (19) gives that the number of iterations for x_T to be $\hat{\varepsilon}$-accurate on $f_S^{(\mu)}$ is

$$T = \tilde{\Omega}_p \left(L^{\frac{2-p}{2\gamma}} R^{\frac{2-p}{\gamma}} n^{\frac{p}{4\gamma}} \hat{\varepsilon}^{-(1-\frac{p}{2})/\gamma} \right)$$

when $\hat{L} \leq p2^{2-p}LR^{2-p}$. Since we need to ensure $\hat{\varepsilon} \leq 6LR^2(\alpha/p)^{1/p}/n^{3/2}$ in order for Theorem 6 to apply, implies we need to have number of iterations as

$$T = \tilde{\Omega}_p \left(n^{\frac{3-p}{2\gamma}} \right)$$

When $n \geq 48p^{2p-4}(p-1)^{2p-4}$, we have that $\mu \leq LR^{1-p}p^{2-p}(p-1)^{p-1}$, which guarantees that $\hat{L} \leq p2^{2-p}R^{2-p}L$ by the developments in Section 9.1 for $p \in (1,2)$.

Consequently, Theorem 6 upper bounds the excess risk of x_T as $\mathbb{E}_S[\delta f(x_T)] \leq 16(\alpha/p)^{1/p} L R^2/n^{1/2}$ for $p \in (1,2)$. □

9.4 Proof of Proposition 12

Proof. Consider the Generalized AGD+ algorithm from (Diakonikolas and Guzmán, 2024) for minimizing $f_S^{(\mu)} = f_S(x) + \mu\psi(x)$, where $f_S(x) : \mathcal{X} \to \mathbb{R}$ is L-smooth, $\mu\psi(x)$ is (μ, p)-uniformly convex, and we assume that $L^{p/2} \geq \mu$. From (Diakonikolas and Guzmán, 2024, (eq. 15)), we know the algorithm takes

$$\begin{aligned} T &\geq c \left(\min \left\{ \left(\frac{1}{\hat{\varepsilon}}\right)^{\frac{p-2}{p+2}} \left(\frac{L^{p/2}}{\mu}\right)^{\frac{2}{p+2}} \log\left(\frac{LD}{\hat{\varepsilon}}\right), \left(\frac{L}{\hat{\varepsilon}}\right)^{\frac{p}{p+2}} \left(\frac{D^{p/2}}{\mu}\right)^{\frac{2}{p+2}} \right\} \right) \\ &\left(\leq c \left(\frac{1}{\hat{\varepsilon}}\right)^{\frac{p-2}{p+2}} \left(\frac{L^{p/2}}{\mu}\right)^{\frac{2}{p+2}} \log\left(\frac{LD^p}{\hat{\varepsilon}}\right) \right) \end{aligned} \tag{22}$$

gradient oracle calls for some constant $c > 0$ in order to achieve $f^{(\mu)}(x_T) - f^{(\mu)}(x_\mu^*) \leq \hat{\varepsilon}$. Due to Lemma 16 we have that $D^p \geq D_\Psi(x_\mu^*, x_0)$ and for the second inequality we picked the first argument of the minimum. *(Here, in the notation of (Diakonikolas and Guzmán, 2024), we take $\phi(u) = \frac{1}{\mu}D_\Psi(u, x_0)$ and upper bound it with D/μ.)*

In order to apply Lemma 5 with $G = 2LR$ and $D = (\alpha/p)^{1/p}R$, we need the Generalized AGD+ algorithm to reach $\hat{\varepsilon}$-minimizer, satisfying both: $\hat{\varepsilon} \leq \mu(\alpha/p)R^p$ and $\hat{\varepsilon} \leq (p/\mu)^{1/(p-1)} \left(\frac{4LR}{n}\right)^{p/(p-1)}$.

When $\mu \leq \alpha^{\frac{1-p}{p}} R^{2-p} \frac{4pL}{n}$, we have that $\mu\frac{\alpha}{p}R^p \leq (p/\mu)^{1/(p-1)} (4LR/n)^{p/(p-1)}$. Then the accuracy we need is $\hat{\varepsilon} \leq \mu\frac{\alpha}{p}R^p$ and the number of required iterations is lower bounded as

$$T \geq c \left(\frac{1}{\mu}\right)^{\frac{p}{p+2}} \left(\frac{\alpha}{p}\right)^{\frac{p-2}{p+2}} R^{-p\frac{p-2}{p+2}} L^{\frac{p}{p+2}} \log\left(\frac{pL}{\mu}\right) \quad . \tag{23}$$

For a given T, we can choose

$$\mu = \left(\frac{c}{T}\right)^{1+\frac{2}{p}} \left(\frac{p}{\alpha}\right)^{1-\frac{2}{p}} R^{2-p} L \left(\log\left(\left(\frac{T}{c}\right)^{1+2/p} p^{2/p} R^{p-2}\right)\right)^{1+\frac{2}{p}}, \tag{24}$$

which for $T \geq ce^{\frac{p}{p+2}} (p/\alpha)^{-2/(p+2)} R^{-p\frac{p-2}{p+2}}$, satisfies (23).

The required bound on $\mu \leq \alpha^{\frac{1-p}{p}} R^{2-p} \frac{4pL}{n}$ is satisfied when

$$T \geq c \left(\frac{n}{4}\right)^{\frac{p}{p+2}} \left(\frac{\alpha}{p}\right)^{\frac{2}{p+2}} \log\left(\left(\frac{T}{c}\right)^{1+\frac{2}{p}} p^{2/p} R^{p-2}\right). \tag{25}$$

By (Baudry et al., 2024, Lemma 4), we have that for $x \geq 3$ and constants $A, B > 0$, if $x \geq 3\frac{A}{B}\log(A)$, then also $x \geq A\log(Bx)$. Applying the result to (25), yields that the Generalized AGD+ algorithm reaches the required accuracy when the number of iterations lower bounded as

$$\begin{aligned} T &\geq 3p^{-\frac{4}{p+2}} \left(1 + \frac{2}{p}\right) c^2 \left(\frac{n}{4}\right)^{\frac{p}{p+2}} R^{2-p} \log\left(c\left(\frac{n}{4}\right)^{\frac{p}{p+2}} p^{-\frac{2}{p+2}}\right) \\ &= \tilde{\Omega}_p\left(c^2 R^{2-p} n^{\frac{p}{p+2}}\right). \end{aligned}$$

For this range of T and the choice of μ in (24), the optimization error is

$$\begin{aligned} \mathcal{E}_{\text{opt}} = f_S(x_T) - f_S(x^*) \leq 2\alpha\mu R^p/p &= 2^{p-1} \left(\frac{c}{T}\right)^{1+\frac{2}{p}} p^{-\frac{2}{p}} D^2 L \left(\log\left(\left(\frac{T}{c}\right)^{1+2/p} p^{2/p} D^{p-2}\right)\right)^{1+\frac{2}{p}} \\ &= \tilde{\mathcal{O}}_p\left(D^2 L \left(\frac{1}{T}\right)^{1+2/p}\right) \end{aligned}$$

and the stability is

$$\begin{aligned} \mathcal{E}_{\text{stab}}(\mathcal{A}) &\leq 3\left(\frac{2p}{n\mu} G^p\right)^{1/(p-1)} \\ &= 3\left(\frac{2^{1+p} p^{2/p} L^p D^p}{LnD^{2-p}} \left(\frac{T}{c}\right)^{1+2/p}\right)^{\frac{1}{p-1}} \left(\log\left(\left(\frac{T}{c}\right)^{1+2/p} D^{p-2}\right)\right)^{-(1+\frac{2}{p})\frac{1}{p-1}} \\ &= \tilde{\mathcal{O}}_p\left(\left(\frac{T^{1+2/p}}{n}\right)^{1/(p-1)} LD^2\right). \end{aligned}$$

□

9.5 Proof of Corollary 13

Proof. For $p \geq 2$, choose $\mu = (\frac{3}{p-1})^{1-\frac{1}{p}} 2^{2/p} p^{1+\frac{1}{p(p-1)}} \alpha^{\frac{1-p}{p}} \left(\frac{1}{n}\right)^{\frac{1}{p}} R^{2-p} L$ as in Theorem 6, and plug it into the convergence rate of the Generalized AGD+ given in (22) in the proof in Section 9.4. We get that the number of required iterations for $\hat{\varepsilon}$ accuracy is

$$T \geq c\left(\frac{p-1}{3}\right)^{2\frac{p-1}{p+2}} 2^{-\frac{4}{p(p+2)}} p^{-\frac{2}{p+2}} n^{\frac{2}{p(p+2)}} R^{2\frac{p-2}{p+2}} L^{\frac{p-2}{p+2}} \left(\frac{1}{\hat{\varepsilon}}\right)^{\frac{p-2}{p+2}} \log\left(\frac{LR^p}{\hat{\varepsilon}}\right),$$

and to achieve accuracy $\hat{\varepsilon} \leq 12LR^2/n^{1+\frac{1}{p}}$, we need

$$T = \tilde{\Omega}_p\left(n^{\frac{p-1}{p+2}}\right)$$

Consequently, Theorem 6 upper bounds the excess risk of x_T as $\mathbb{E}_S[\delta f(x_T)] \leq 16LR^2/n^{1/p}$ for $p \geq 2$. □