Appendix I: Simulation Plots

In Section 5 we discuss our simulations on ℓ_p unit balls. The hyperparameters are the exponent p, dimension d, parameter r that determines the expected cost vector, and noise types (1) and (2). For $d = 32$, $p = 1.5$ and a range of r-values Figure 1 shows the worst performance for noise types $(1), (2)$ occurs for $r = 0, 8$ respectively. Figure 3 considers both worst cases and varies the dimension.

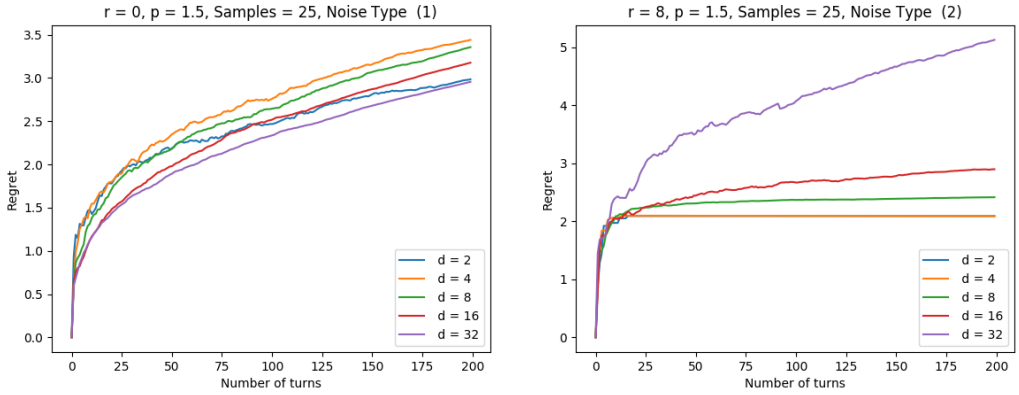

Figure 3: Online Lazy Gradient Descent on the $\ell_{1.5}$ unit ball, varying dimension.

Figure 3 shows the performance improves slightly in higher dimension for noise type (1) and degrades in higher dimension for noise type (2). Figure 4 varies the exponent p for both worst cases.

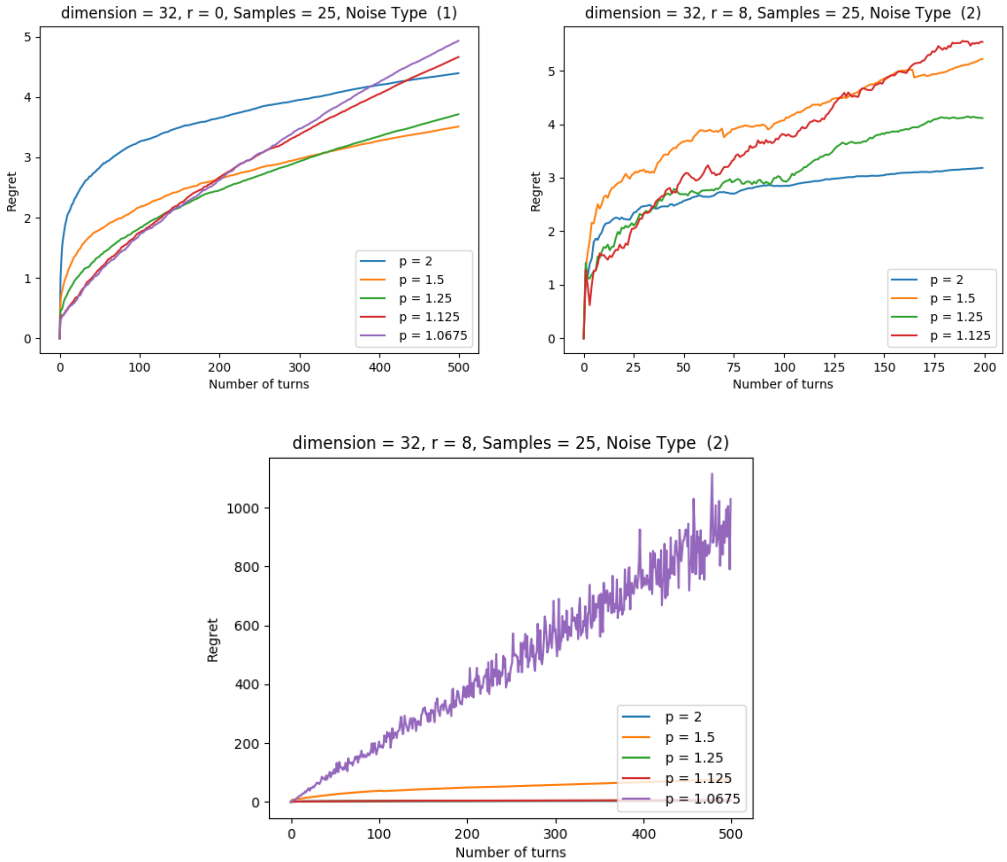

Figure 4: Online Lazy Gradient Descent Descent on the ℓ_p unit ball, varying p.

Figure 4 shows that in all cases the performance degrades as $p \to 1$; though for noise type (1) it also takes longer for the worse performance to emerge. This is unsurprising since the strong convexity parameter $m = (p-1)d^{\frac{1}{2}-\frac{1}{p}}$ of the ℓ_p unit ball (see Corollary 2 of [17]) tends to 0 as $p \to 1$. Hence the $O\big(\frac{L^2}{m}\log N\big)$ bound in Theorems 1 and 3 also tends to infinity.

Next we present additional plots for our simulations on Schatten matrix norm unit balls. Again the hyperparameters are the exponent p, dimension d, parameter r that determines the expected cost vector, and noise types (1),(2) and (3). For $d = 5$, $p = 1.5$ and a range of r-values Figure 2(a) shows the worst performance for noise types (1) occurs for $r = 0$. Figure 5 shows the different noise types have virtually identical behaviour.

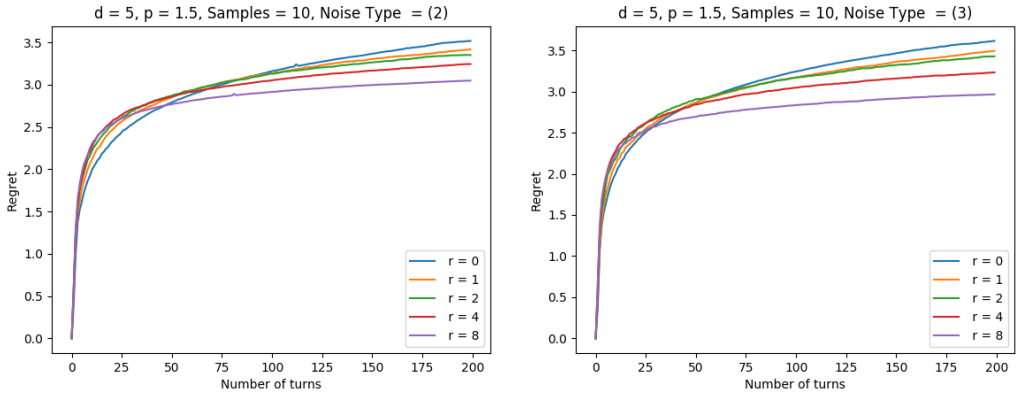

Figure 5: Online Lazy Gradient Descent on the Schatten ball $B_p(1)$, varying cost vectors.

Figure 6 varies the dimension and parameter p. Small dimension seems to give slightly better performance. Similar to before the performance degrades as $p \to 1$, though the influence of p is less significant than in Figure 4 for ℓ_p balls.

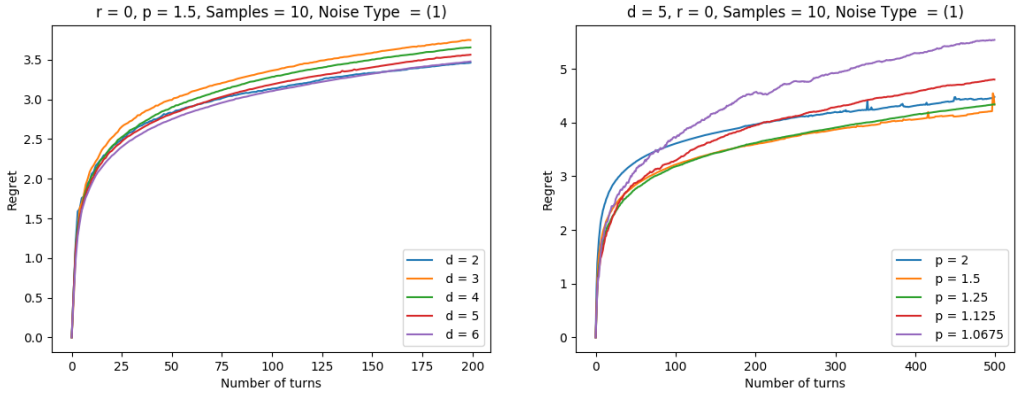

Figure 6: Online Lazy Gradient Descent on the Schatten ball $B_p(1)$ varying p and dimension.

Finally we prove Lemma 13 to justify only considering the expectation to be a diagonal matrix.

Lemma 13. Suppose $a_1, a_2, \ldots \in \mathbb{R}^{d\times d}$ are i.i.d cost vectors with $\mathbb{E}[a_n] = a$. There is another set of i.i.d cost vectors $b_1, b_2, \ldots$ such that $\mathbb{E}[b_n] = b$ is diagonal and $\|a_n - a\| = \|b_n - b\|$; $\|a_n\| = \|b_n\|$; $\|a\| = \|b\|$ and the expected regret of playing Lazy Gradient Descent on $B_p(1)$ against $a_1, a_2, \ldots$ with $x_0 = 0$ is the same as playing against $b_1, b_2, \ldots$ with $x_0 = 0$.

Proof. Let $X \subset \mathbb{R}^d$ be an arbitrary domain, c_1, c_2, be cost vectors and $L : \mathbb{R}^d \to \mathbb{R}^d$ an orthogonal transformation with $L(X) = X$. The actions of Lazy Gradient Descent with $x_0 = 0$ against c_n are $x_n = \operatorname{argmin}_{x\in X} \left\|x - \frac{\eta}{\sqrt{n}}\sum_{i=1}^n c_i\right\|$. Since L is an isometry we have $x_n = \operatorname{argmin}_{x\in X} \left\|L\big(x - \frac{\eta}{\sqrt{n}}\sum_{i=1}^n c_i\big)\right\| = \operatorname{argmin}_{x\in X} \left\|Lx - \frac{\eta}{\sqrt{n}}\sum_{i=1}^n \widetilde{c}_i\right\|$ for $\widetilde{c}_i = Lc_i$. The minimum is achieved when

$Lx = \widetilde{x}_n$ where $\widetilde{x}_n$ are the actions of Lazy Gradient Descent with $x_0 = 0$ against $\widetilde{c}_n$. Hence we have $x_n = L^{-1}\widetilde{x}_n = L^*\widetilde{x}_n$. Similarly each $x^* \in \arg\min_{x\in X} \sum_{i=1}^N c_i \cdot x$ has the form $L^*\widetilde{x}^*$ for some $\widetilde{x}^* \in \arg\min_{x\in X} \sum_{i=1}^N \widetilde{c}_i \cdot x$. Hence the regret against c_n can be written $R_N = \sum_{i=1}^N c_i \cdot (x_i - x^*) = \sum_{i=1}^N c_i \cdot L^*(\widetilde{x}_i - \widetilde{x}^*) = \sum_{i=1}^N (Lc_i) \cdot (\widetilde{x}_i - \widetilde{x}^*) = \sum_{i=1}^N \widetilde{c}_i \cdot (\widetilde{x}_i - \widetilde{x}^*)$ and we see the regret for playing against c_n is the same as playing against $\widetilde{c}_n$.

To define b_n let $\mathbb{E}[a_n] = a = U\Sigma V$ be the singular value decomposition and $b_n = U^* a_n V^*$. Then $\mathbb{E}[b_n] = U^*\mathbb{E}[a_n]V^* = U^*U\Sigma VV^* = \Sigma$ is diagonal as required.

We claim the linear operator $L : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$; $A \mapsto U^*AV^*$ is orthogonal with respect to the inner product $A \bullet B = \sum_{i,j} A_{ij}B_{ij}$. It is enough to show $A \mapsto AV^*$ and $A \mapsto U^*A$ or equivalently the inverses $A \mapsto AV$ and $A \mapsto UA$ are orthogonal. For write.

$$\begin{aligned} UA \bullet B &= \sum_{i,j} (UA)_{ij}B_{ij} = \sum_{i,j} \Big(\sum_k U_{ik}A_{kj} \Big) B_{ij} = \sum_{i,j,k} A_{kj}U_{ik}B_{ij} = \sum_{i,j,k} A_{kj}U^*_{ki}B_{ij} \\ &= \sum_{j,k} A_{kj} \Big(\sum_i U^*_{ki}B_{ij} \Big) = \sum_{j,k} A_{kj}(U^*B)_{kj} = \sum_{i,j} A_{ij}(U^*B)_{ij} = A \bullet (U^*B) \end{aligned}$$

Hence the adjoint of $A \mapsto UA$ is $A \mapsto U^*B$. Since U is orthogonal this is also the inverse and so $A \mapsto UA$ is orthogonal. The proof for $A \mapsto AV$ is similar. Hence by composition L is orthogonal. Thus it is an isometry and we have $\|b_n - b\| = \|La_n - La\| = \|L(a_n - a)\| = \|a_n - a\|$ and likewise $\|a_n\| = \|b_n\|$ and $\|a\| = \|b\|$.

Once we have shown L preserves $B_p(1)$, the first part of the proof says playing Lazy Gradient Descent against $a_1, a_2 \dots$ gives the same regret as playing against $b_1, b_2 \dots$. To see L preserves $B_p(1)$ observe for any A with singular value decomposition $A = U_1\Sigma_1V_1$ that LA has decomposition $A = U_2\Sigma_1V_2$ for $U_2 = U^*U_1$ and $V_2 = V_1V^*$. In particular the singular values are unchanged and $\|LA\|_{S(p)} = \|A\|_{S(p)}$. □

Appendix II: Differential Geometry

Our proof of Theorem 1 focuses on how the actions x_n approach the expected minimiser x^* in terms of both distance and direction. By Assumption 1 the boundary is a smooth surface, and this motivates our use of differential geometry results. For comparison the existing literature focuses on the costs rather than the actions, and presents no reason to focus on the boundary surface.

Strong convexity of the domain is encoded in the differential geometry of the boundary (See Proposition 4 of [20]). For a one dimension function $f : \mathbb{R} \to \mathbb{R}$ with second derivative f'' we have m-strong convexity if and only if $f''(x) \geq m$ for all $x \in \mathbb{R}$. The second derivative can be thought of as the rate of change of tangent vector or equivalently as the rate of change of normal vector to the graph.

For an implicitly defined surface $\mathcal{M} = \{x \in \mathbb{R}^d : F(x_1, \dots, x_d) = 0\}$ the first and second derivatives $f'(x)$ and $f''(x)$ correspond to the unit normal vector $N(x)$ and its rate of change $\nabla N(x)$. Note since we can write $N(x) = \nabla F(x)/\|\nabla F(x)\|$ it makes sense to speak of the normal operator defined on the whole space rather than just the surface. Since the unit normal is vector-valued the rate of change is a $d \times d$ matrix. This matrix gives a linear operator $\mathbb{R}^d \to \mathbb{R}^d$ via matrix multiplication.

Since we cannot move perpendicular to the normal vector without leaving the surface, we must restrict this linear operator to the $(d-1)$-dimensional tangent space $T_x\mathcal{M}$ to the surface at x, in order to talk about how the normal changes as we move across the manifold. Put another way, there are different choices of the function F that give the same surface $\mathcal{M} = \{x \in \mathbb{R}^d : F(x_1, \dots, x_d) = 0\}$. For example $2F$ and F^2 have the same zero set. Hence the object $N(x) = \nabla F(x)/\|\nabla F(x)\|$ contains information about the choice of function as well as the surface. By ignoring the action on the normal direction we keep only the information about the surface.

The restriction of the operator is called the differential $DN(x)$ or **shape operator** of the surface at x. It is useful to think of $\nabla N(x)$ and $DN(x)$ as linear operators rather than matrices, since the standard basis on $\mathbb{R}^d$ does not naturally give a basis for $T_x\mathcal{M}$ or matrix form for $DN(x)$. However we can still talk about the eigenvalues and vectors since they are basis-independent objects. We have m-strong convexity of $\mathcal{M}$ if and only if $\lambda_1, \dots \lambda_{d-1} \geq m$ for the eigenvalues $\lambda_1, \dots \lambda_{d-1}$ of $DN(x)$.

Of course eigenvalues of an operator are only defined if the domain and range are the same space. Hence to talk about eigenvalues of $DN(x) : T_x\mathcal{M} \to T_x\mathcal{M}$ we must know $\nabla N(x)$ transforms elements of $T_x\mathcal{M}$ into elements of the same subspace. To see this recall each tangent vector $v \in T_x\mathcal{M}$ is witnessed by some path $\gamma : (-\varepsilon, \varepsilon) \to \mathcal{M}$ with $\gamma(0) = x$ and $\gamma'(0) = v$. Hence by the chain rule we have we have $\nabla N(x)v = \nabla N(x)\gamma'(0) = \frac{d}{dt}\big|_{t=0} N(\gamma(t))$. Let $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ be the $(d-1)$-dimensional unit sphere, and write the right-hand-side of the above as $\Gamma'(0)$ for the path $\Gamma : (-\varepsilon, \varepsilon) \to \mathbb{S}^{d-1}$ with $\Gamma(0) = N(x)$ and $\Gamma(t) = N(\gamma(t))$. Note Γ is well-defined since each unit normal $N(\gamma(t))$ is an element of the unit sphere. Thus the right-hand-side is tangent to the sphere at the point $\Gamma(0) = N(x)$. Hence we have shown $\nabla N(x)v \in T_{N(x)}\mathbb{S}^{d-1}$. But the tangent space to the sphere at $N(x)$ is just the orthogonal complement of $\{N(x)\}$ which is also the tangent space $T_x\mathcal{M}$. Hence we have $\nabla N(x)v \in T_x\mathcal{M}$ as required. We conclude it makes sense to speak of the eigenvalues of $DN(x) : T_x\mathcal{M} \to T_x\mathcal{M}$.

One non-obvious fact about the shape operator $DN(x)$ is that it is symmetric. This holds even if the non-restricted operator $\nabla N(x)$ fails to be symmetric.

The intuition behind the symmetry is to choose coordinates (or equivalently perform a rotation and translation) so $\mathcal{M}$ is locally represented as the graph $\{x \in \mathbb{R}^d : x_d = G(x_1, \ldots, x_{d-1})\}$ of a twice differentiable function G with $G(x) = 0$ and $\nabla G(x) = e_d$ pointing downwards. Writing $(a; b) = (a_1, \ldots a_r, b_1, \ldots, b_l)$ we see the normal direction at $(x; G(x))$ points along $(\nabla G(x); 1)$. Hence for $j = 1, 2, \ldots, d-1$ we have

$$\begin{aligned}\nabla N(0)e_j &= \lim_{n\to\infty} \frac{N(e_j/n) - N(0)}{n} = \lim_{n\to\infty} \frac{1}{n}\left(\frac{(\nabla G(e_j/n); 1)}{\sqrt{\nabla G(e_j/n)^2 + 1}} - (0; 1)\right) \\ &= \lim_{n\to\infty} \frac{(\nabla G(e_j/n); 1) - (0; 1)}{n} = \lim_{n\to\infty} \frac{(\nabla G(e_j/n); 0)}{n} = (\nabla^2 G(0)e_j; 0)\end{aligned}$$

for $\nabla^2 G(0)$ the Hessian matrix of second partial derivatives of G at 0. Hence the the restriction of $\nabla N(0)$ to the tangent space is the same as the Hessian, and the shape operator $DN(0)$ is symmetric.

Since the shape operator is in general symmetric, standard linear algebra then says $DN(x)$ has eigenvectors $v_1, \ldots, v_{d-1}$ with eigenvalues $\lambda_1, \ldots, \lambda_{d-1}$ such that each $e_i^T DN(x)e_j = \lambda_i \delta_{ij}$. In other words the matrix representation is $\text{diag}(\lambda_1, \ldots, \lambda_{d-1})$ and we see strong convexity is equivalent to having $v^T DN(x)v \geq m$ for all $v \in T_x\mathcal{M}$.

Curvature

Here we give a formal treatment of the notion of curvature of a manifold, as outlined in the previous subsection. Recall the domain $X \subset \mathbb{R}^d$ is assumed to be m-strongly convex and the boundary $\mathcal{M} = \partial X$ is a $(d-1)$-dimensional C^2 manifold. Namely each $z \in \mathcal{M}$ has a neighborhood U in $\mathbb{R}^d$ and C^2 function $F : U \to \mathbb{R}$ with nonzero gradient such that $\mathcal{M} \cap U = \{x \in U : F(x) = 0\}$. Such a function is called a **coordinate patch** at z. Note this definition is slightly different to that in the literature. For equivalent definitions of C^2 manifolds see [38] Theorem 1.41.

Much of the following discussion is standard. However we were unable to find a good reference for differential geometry of hypersurfaces in $\mathbb{R}^d$. Much of the field instead focuses on properties that are intrinsic to the manifold itself and not the particular embedding in $\mathbb{R}^d$. First we recall some standard machinery from differential geometry.

We write $N(z)$ for the unit outwards normal at $z \in \mathcal{M}$. For each $z \in \mathcal{M}$ write $T_x\mathcal{M} = \{v \in \mathbb{R}^d : v \perp N(z)\}$ for the tangent space at z to $\mathcal{M}$. The map $\phi : \mathcal{M} \to \mathcal{N}$ between manifolds is said to be differentiable at z to mean there exists a linear operator $D\phi(z) : T_x\mathcal{M} \to T_{\phi(z)}\mathcal{N}$ with the following property: Given any $v \in T_x\mathcal{M}$ and differentiable path $\gamma : (-\varepsilon, \varepsilon) \to \mathcal{M}$ with $\gamma(0) = z$ and $\gamma'(0) = v$ we have $D\phi(z)v = \frac{d}{dt}\big|_{t=0} \phi(\gamma(t))$.

Let $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ be the unit sphere and $N : \mathcal{M} \to \mathbb{S}^{d-1}$ the unit outwards normal. Since each point of $\mathcal{M}$ is contained in some coordinate patch F we can represent N locally as $N(x) = \frac{\nabla F(x)}{\|\nabla F(x)\|}$. Note by definition $\nabla F \neq 0$ over U hence the expression is well-defined. Since F is C^2 it follows N is locally continuously differentiable. This is the same as being globally continuously differentiable.

In the previous subsection we have already shown the differential $DN(x) : T_x\mathcal{M} \to T_x\mathcal{M}$ is well-defined. Recall the definition of the curvature.

Definition D1. The linear operator $\nabla N(x) : \mathbb{R}^d \to \mathbb{R}^d$ restricts to an operator $DN(x) : T_x\mathcal{M} \to T_x\mathcal{M}$. The curvature m_x of $\mathcal{M}$ at x is defined as the smallest eigenvalue of $DN(x)$.

In the previous subsection we gave an example of why $DN(x)$ is symmetric. The formal proof is contained in Lemma D2.

Lemma D2. For each $x \in \mathcal{M}$ there is a scalar function $G : \mathbb{R}^d \times T_x\mathcal{M}$ such that each

$$DN(x)v = \frac{\nabla^2 F(x)}{\|\nabla F(x)\|}v - \nabla F(x)G(x,v)$$

Moreover for any $w, v \in T_x\mathcal{M}$ we have $w^T \nabla N(x)v = w^T \frac{\nabla^2 F(x)}{\|\nabla F(x)\|}v$.

Proof. By choosing a coordinate patch F at x we can write $N(y) = F(y)/\|\nabla F(y)\|$ and so

$$\begin{aligned} DN(x)v &= \lim_{t\to 0} \frac{1}{t}\left(\frac{\nabla F(x+tv)}{\|\nabla F(x+tv)\|} - \frac{\nabla F(x)}{\|\nabla F(x)\|}\right) \\ &= \lim_{t\to 0} \frac{1}{t}\left(\frac{\nabla F(x+tv)}{\|\nabla F(x+tv)\|} - \frac{\nabla F(x+tv)}{\|\nabla F(x)\|} + \frac{\nabla F(x+tv)}{\|\nabla F(x)\|} - \frac{\nabla F(x)}{\|\nabla F(x)\|}\right) \\ = \nabla F(x) \lim_{t\to 0} \frac{1}{t}\left(\frac{1}{\|\nabla F(x+tv)\|} - \frac{1}{\|\nabla F(x)\|}\right) &+ \frac{1}{\|\nabla F(x)\|} \lim_{t\to 0}\left(\frac{\nabla F(x+tv) - \nabla F(x)}{t}\right) \end{aligned}$$

The second term is $\frac{\nabla^2 F(x)}{\|\nabla F(x)\|}v$ and the first is $\nabla F(x)\frac{d}{dt}\big|_{t=0}\frac{1}{\|\nabla F(x+tv)\|}$. Since $\nabla F(x) \neq 0$ the derivative is some well-defined scalar function $G(x,v)$. This proves the first part of the claim. For the second part write

$$w^T DN(x)v = \frac{w^T \nabla^2 F(x)v}{\|\nabla F(x)\|} - w^T \nabla F(x)G(x,v).$$

Since w is a tangent vector it is orthogonal to the normal and the second term vanishes. □

Proposition 4 of [20] says that $\mathcal{M} = \partial X$ having curvature at least $m > 0$ at each point is equivalent to X being m-strongly convex. Since $m > 0$ all the eigenvalues of $\nabla N(x)$ and $DN(x)$ are positive and we have the next lemma.

Lemma 2. For each $z \in \mathcal{M}$ and unit vector v tangent to $\mathcal{M}$ at z we have $\|\nabla N(z)v\| \geq m$.

Proof. Since $v \in T_z\mathcal{M}$ we have $N(z)v = DN(z)v$. The second part of Lemma D2 says $DN(z) : T_z\mathcal{M} \to T_z\mathcal{M}$ is a symmetric operator. Hence $T_z\mathcal{M}$ has an orthonormal basis of eigenvectors $u_1, \ldots, u_{d-1}$ with eigenvalues $\lambda_1, \ldots, \lambda_{d-1}$. Hence we have $v = \sum_{j=1}^{d-1} \alpha_j u_j$ for some $\alpha_i \in \mathbb{R}$ such that $\sum_{j=1}^{d-1} \alpha_j^2 = 1$. It follows $DN(z)v = \sum_{j=1}^{d-1} \alpha_j \lambda_j u_j$ and so $\|DN(z)v\|^2 = \sum_{j=1}^{d-1} \alpha_j^2 \lambda_j^2$. The right-hand-side is minimised for $\alpha_j = 1$ for $\lambda_j = \min_j \lambda_j^2$ and all other $\alpha_k = 0$. Hence we have $\|DN(z)v\|^2 \geq \min \lambda_j^2$. Since we assume positive curvature all eigenvalues are positive and so $\|DN(z)v\| \geq \min_j |\lambda_j| = \min_j \lambda_j = m$ as required. □

Lemma 3. For each coordinate patch F at $z \in \mathcal{M}$ and each unit vector tangent v to $\mathcal{M}$ at z we have $\frac{v^T \nabla^2 F(z)v}{\|\nabla F(z)\|} \geq m$.

Proof. Lemma D2 says $\frac{v^T \nabla^2 F(z)v}{\|\nabla F(z)\|} = v^T DN(z)v$. For u_j, λ_j, α_j from the proof of Lemma 2 we have $v^T DN(z)v = \left(\sum_{j=1}^{d-1} \alpha_j u_j\right)^T DN(z) \sum_{j=1}^{d-1} \alpha_j u_j = \left(\sum_{j=1}^{d-1} \alpha_j u_j\right)^T \sum_{j=1}^{d-1} \alpha_j \lambda_j u_j = \sum_{j=1}^{d-1} \alpha_j^2 \lambda_j$ since u_j are orthonormal. Since we assume positive curvature all $\lambda_j \geq 0$ and the right-hand-side is minimised same as the proof of Lemma 2. □

Next we define a counterpart to the curvature. While the curvature lower bounds how quickly the unit normal changes as we vary the basepoint, the counterpart upper bounds the change.

Lemma 4. There exist constants $M, r > 0$ such that each $z \in \mathcal{M}$ has a coordinate patch $F : B(z,r) \to \mathbb{R}$ with $\|\nabla N(y)\|, \frac{\|\nabla^2 F(y)\|}{\|\nabla F(y)\|} \leq M$ for all $y \in B(z,r)$.

Proof. For each $x \in \mathcal{M}$ let $F_x : U_x \to \mathbb{R}$ be a coordinate patch at x. For each x let $V_x \subset U_x$ be a compact neighborhood of x chosen small enough that $\nabla F_x \neq 0$ over V_x. Since F_x is C^2 we see $\frac{\|\nabla^2 F_x(y)\|}{\|\nabla F_x(y)\|}$ is continuous over V_x hence bounded by some M_x. Likewise the function $y \mapsto \nabla N(y)$ that gives the matrix of partial derivatives of $\frac{\nabla F_x(y)}{\|\nabla F_x(y)\|}$ is continuous. Hence $y \mapsto \|\nabla N(y)\|$ is bounded over V_x by some M'_x.

Since $\mathcal{M}$ is compact it is covered by some finite subcollection $V^\circ_{x(1)}, \ldots, V^\circ_{x(n)}$. The Lebesgue covering theorem, see [15] Theorem 4.3.31, gives $r > 0$ such that for each $x \in \mathcal{M}$ the ball $B(x,r)$ is contained in some $V^\circ_{x(i)}$. Hence we can take F as the restriction of $F_{x(i)}$ and the constant $M = \max\{M_{x(i)}, M'_{x(i)} : i \leq n\}$ □

Remark: The constants M and r in Lemma 4 depend only on the geometry of $\mathcal{M}$ and not on the manner $\mathcal{M}$ is embedded in $\mathbb{R}^d$. To see this recall the operator norms $\|\nabla N(y)\|, \|\nabla^2 F(y)\|$ and $\|\nabla F(y)\|$ are unchanged by translations and rotations. Hence if the constants satisfy the conditions for one embedding we can transform between embeddings to show they satisfy for all other embeddings. However it is not obvious how to express M and r in terms of intrinsic properties of $\mathcal{M}$. For r we suspect the tubular neighborhood theorem [36] can be used to show a single coordinate patch is sufficient. However it is difficult to find references for such theorems for embedded C^2 manifolds rather than C^∞ manifolds. For M we suspect the coordinate patch F can be selected so that for $N(x) = \nabla F(x)/\|\nabla F(x)\|$ we have $\nabla N(x)$ coincides with $DN(x)$ on the tangent space and vanishes on the normal direction, and so M is the largest eigenvalue of the shape operator.

Main Result

Here we prove our main differential geometry result. Our Proposition 1 gives a lower bound for the change of unit normal, as we vary the basepoint along the boundary surface of a C^2 and m-strongly convex domain.

For a one-dimensional example suppose $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable. That means for any $z \in \mathbb{R}$ the derivative has linear approximation $f'(y) = f'(z) + f''(z)(y-z) + o(\|y-z\|)$. Hence z has a neighborhood U with for example $\|f'(y) - f'(z)\| \geq \frac{|f''(z)|}{2}\|y-z\|$ for all $y \in U$. In case f is m-strongly convex we have moreover $\|f'(y) - f'(z)\| \geq \frac{m}{2}\|y-z\|$ for all $y, z \in \mathbb{R}$.

Proposition 1 is analogous to the above except (a) the right-hand-side uses the geodesic rather than norm distance and (b) the coefficient $\|\nabla N(z)v\|$ mentions the direction v from x to y as well as the derivative $\nabla N(z)$ of the normal vector. For example consider the graph of a quadratic $z = \alpha x^2 + \beta y^2$ for $\alpha, \beta \geq 0$. For perturbations from 0 along the x, y-axis we can bound the change in normal by the coefficients α, β respectively.

The closest existing result seems to be Vial, 1982 Theorem 1(v) [35] which says $\|N(y) - N(z)\| \geq m\|x-y\|$ for all $y, z \in \mathcal{M}$. This result is superior in that it applies globally rather than in a neighborhood of z. However the theorem is insufficient here as it does not mention the direction. This is important because our main proof uses the form $d(y,z) \leq \frac{2}{\|\nabla N(z)v\|}\|N(y) - N(z)\|$ of Proposition 1 to upper bound the expression $E = \|\nabla N(z)v\| d(y,z)^2$. The factors of $\|\nabla N(z)v\|$ cancel and we get $E \leq \frac{4}{\|\nabla N(z)v\|}\|N(y) - N(z)\|^2$ which is at most $\frac{4}{m}\|N(y) - N(z)\|^2$ by Lemma 2. This ultimately gives an $O\big(\frac{L^2}{m}\log N\big)$ bound for pseudo-regret. If we instead try to use the Vial theorem we only get $E \leq \frac{\|N(z)v\|}{m^2}\|N(y) - N(z)\|^2$ and an $O\big(\frac{L^2}{m}\frac{\overline{m}}{m}\log N\big)$ bound for $\overline{m}$ the largest eigenvalue of the shape operator. Hence our new result is needed to get the optimal coefficient.

Lemma 5. Let M be the constant from Lemma 4. For each geodesic $\gamma : [0, d(x,y)] \to \mathcal{M}$ from x to y we have for $d = d(x,y)$ and all $t \leq d$ the inequalities

$$(1)\ \|\gamma''(t)\| \leq M \qquad (2)\ \|\gamma(t) - \gamma(0) - t\gamma'(0)\| \leq (M/2)t^2 \qquad (3)\ \big|\|y-x\| - d\big| \leq (M/2)d^2$$

Proof. To prove (1) let F be a coordinate patch from Lemma 4 and differentiate $F(\gamma(t)) = 0$ twice to get $\nabla^2 F(\gamma(t))\gamma'(t) + \nabla F(\gamma(t))\gamma''(t) = 0$ and so $\|\nabla F(\gamma(t))\gamma''(t)\| = \|\nabla^2 F(\gamma(t))\gamma'(t)\|$. Since γ is a geodesic [29] says $\gamma''(t)$ is normal to the surface and the left-hand-side is $\|\nabla F(\gamma(t))\gamma''(t)\| = \|\nabla F(\gamma(t))\|\|\gamma''(t)\|$. Divide to get

$$\|\gamma''(t)\| = \left\| \frac{\nabla^2 F(\gamma(t))}{\|\nabla F(\gamma(t))\|} \gamma'(t) \right\| \leq M\|\gamma'(t)\| \leq M$$

since γ is a unit-speed path. For (2) write

$$\gamma(t) - \gamma(0) = \int_0^t \gamma'(x)dx = \int_0^t \left(\gamma'(0) + \int_0^s \gamma''(s)ds \right) dx = t\gamma'(0) + \int_0^t \int_0^s \gamma''(s)dsdx$$
$$\implies \|\gamma(t) - \gamma(0) - t\gamma'(0)\| \leq \int_0^t \int_0^s \|\gamma''(s)\|dsdx \leq M \int_0^t \int_0^s dsdx = \frac{M}{2}t^2$$

To get (3) recall $x = \gamma(0)$ and $y = \gamma(d)$. Since γ has unit speed the triangle inequality gives

$$\big|\|y - x\| - d\big| = \big|\|\gamma(d) - \gamma(0)\| - d\|\gamma'(0)\|\big| \leq \|\gamma(d) - \gamma(0) - d\gamma'(0)\| \leq \frac{M}{2}d^2$$

where the last inequality uses (2). □

Proposition 1. Each $z \in \mathcal{M}$ has a neighborhood U in $\mathbb{R}^d$ such that for all $y \in U \cap \mathcal{M}$ and direction v from z to y we have

$$\|N(y) - N(z)\| \geq \frac{\|\nabla N(z)v\|}{2} d(y, z)$$

Proof. Recall the normal unit vector can be expressed $N(z) = \frac{\nabla F(z)}{\|\nabla F(z)\|}$ for any coordinate patch F at z. Since F is C^2 the normal vector function is differentiable. Hence z has a neighborhood U such that for all $y \in U \cap \mathcal{M}$ we have

$$\|N(y) - N(z) - \nabla N(z)(y - z)\| \leq \frac{m}{4}\|y - z\|$$
$$\implies \|N(y) - N(z) - d\nabla N(z)v\| \leq \frac{m}{4}\|y - z\| + \|\nabla N(z)((y - z) - dv)\|$$

for $d = d(y, z)$. Shrink U if necessary to have geodesic diameter less than $m/2M^2$ for the constant M from Lemma 4. The reverse triangle inequality applied to the above gives

$$\big|\|N(y) - N(z)\| - d\|\nabla N(z)v\|\big| \leq \frac{m}{4}\|y - z\| + \big\|\nabla N(z)\big((y - z) - dv\big)\big\|$$
$$\implies \big|\|N(y) - N(z)\| - d\|\nabla N(z)v\|\big| \leq \frac{m}{4}d(y, z) + \big\|\nabla N(z)\|\|(y - z) - dv\big\|. \tag{4}$$

where we have used $\|z - y\| \leq d(y, z)$. Use Lemma 5 to bound the rightmost term as

$$\big\|\nabla N(z)\|\|(y - z) - dv\big\| \leq M\|y - z - dv\| \leq \frac{M^2}{2}d(y, z)^2 \leq \frac{M^2}{2}\frac{m}{2M^2}d(y, z) = \frac{m}{4}d(y, z).$$

Hence the right-hand-side of (4) is at most $\frac{m}{2}d(y, z)$ and we get

$$\|N(y) - N(z)\| - \|\nabla N(z)v\|d(y, z) \geq \frac{m}{2}d(y, z)$$
$$\implies \|N(y) - N(z)\| \geq \left(\|\nabla N(z)v\| - \frac{m}{2}\right) d(y, z) \geq \frac{\|\nabla N(z)v\|}{2} d(y, z).$$

where the last line uses Lemma 2. □

Appendix III: Probability

The important concentration inequality is the following corollary to [28] Theorem 3.5.

Proposition 2. Suppose $X_1, X_2, \ldots \in \mathbb{R}^d$ are independent random variables with each $\mathbb{E}[X_n] = 0$ and $\|X_n\| \leq R$. For each $t > 0$ we have

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^{n} X_i\right\| > t\right) \leq 2\exp\left(-\frac{t^2}{2R^2}n\right).$$

To apply the above we first prove some geometric lemmas.

Lemma P1. For any $a, \varepsilon \in \mathbb{R}^d$ with $\|\varepsilon\| \leq \dfrac{\|a\|}{2}$ we have $\left\|\dfrac{a+\varepsilon}{\|a+\varepsilon\|} - \dfrac{a}{\|a\|}\right\| \leq 6\dfrac{\|\varepsilon\|}{\|a\|}$.

Proof. Write $\dfrac{a+\varepsilon}{\|a+\varepsilon\|} = \dfrac{a}{\|a\|} + a\left(\dfrac{1}{\|a+\varepsilon\|} - \dfrac{1}{\|a\|}\right) + \dfrac{\varepsilon}{\|a+\varepsilon\|}$ and so

$$\begin{aligned}\left\|\frac{a+\varepsilon}{\|a+\varepsilon\|} - \frac{a}{\|a\|}\right\| &\leq \frac{\|\varepsilon\|}{\|a+\varepsilon\|} + \|a\|\left|\frac{1}{\|a+\varepsilon\|} - \frac{1}{\|a\|}\right| \\ &\leq \frac{\|\varepsilon\|}{\|a\|/2} + \|a\|\left|\frac{1}{\|a+\varepsilon\|} - \frac{1}{\|a\|}\right| \leq 2\frac{\|\varepsilon\|}{\|a\|} + \|a\|\frac{\|\varepsilon\|}{(\|a\|/2)^2} = 6\frac{\|\varepsilon\|}{\|a\|}\end{aligned}$$

where the second line uses convexity of the reciprocal function. □

Lemma P2. Suppose $n \geq 16D^2/\eta^2\|a\|^2$ and $\left\|\sum_{i=1}^{n}(a_i - a)\right\| \leq n\|a\|/4$. The Lazy Gradient Descent actions $x_{n+1} = \Pi_X\big(-\frac{\eta}{\sqrt{n}}\sum_{i=1}^{n} a_i\big)$ give each

$$\frac{\|a\|}{6}\|N(x^*) - N(x_{n+1})\| \leq \frac{D}{\eta\sqrt{n}} + \frac{1}{n}\left\|\sum_{i=1}^{n}(a_i - a)\right\|. \tag{5}$$

Proof. First we claim $-\frac{\eta}{\sqrt{n}}\sum_{i=1}^{n} a_i - x_{n+1}$ is normal outwards to X at x_{n+1}. To that end recall for the Euclidean projection Π_X and any $y \notin X$ we know $y - \Pi_X(y)$ is outwards normal to X at $\Pi_X(y)$. For $y = -\frac{\eta}{\sqrt{n}}\sum_{i=1}^{n} a_i$ we have $\Pi_X(y) = x_{n+1}$. Since $\|x\| \leq D$ for all $x \in X$ it is enough to show $\|\frac{\eta}{\sqrt{n}}\sum_{i=1}^{n} a_i\| > D$. To that end write $\frac{\eta}{\sqrt{n}}\sum_{i=1}^{n} a_i = \frac{\eta}{\sqrt{n}}\sum_{i=1}^{n}(a_i - a) + \eta\sqrt{n}a$. Hence the reverse triangle inequality gives

$$\left\|\frac{\eta}{\sqrt{n}}\sum_{i=1}^{n} a_i\right\| \geq \left|\|\eta\sqrt{n}a\| - \|\frac{\eta}{\sqrt{n}}\sum_{i=1}^{n}(a_i - a)\|\right| \geq \eta\sqrt{n}\|a\| - \frac{\eta}{\sqrt{n}}\left\|\sum_{i=1}^{n}(a_i - a)\right\|$$

By assumption on $\left\|\sum_{i=1}^{n}(a_i - a)\right\|$ the above is at least

$$\eta\sqrt{n}\|a\| - \frac{\eta}{\sqrt{n}}\frac{n\|a\|}{4} = \left(\eta\|a\| - \eta\frac{\|a\|}{4}\right)\sqrt{n} = \frac{3}{4}\eta\|a\|\sqrt{n}$$

By assumption on n we have $\sqrt{n} \geq 4D/\eta\|a\|$ and the above is at least $3D$. Hence $\|\frac{\eta}{\sqrt{n}}\sum_{i=1}^{n} a_i\| \geq 3D > D$ as required.

Since $x^* = \operatorname{argmin}\{a \cdot x : x \in X\}$ the unit outwards normal at x^* is $N(x^*) = -\frac{a}{\|a\|}$. In the notation of Lemma P1 take $\varepsilon = \frac{1}{n}\sum_{i=1}^{n}(a_i - a) + \frac{x_{n+1}}{\eta\sqrt{n}}$. We claim $\frac{a+\varepsilon}{\|a+\varepsilon\|} = -N(x_{n+1})$. To see this observe $a + \varepsilon = \frac{1}{n}\sum_{i=1}^{n} a_i + \frac{x_{n+1}}{\eta\sqrt{n}} = -\frac{1}{\eta\sqrt{n}}\left(-\frac{\eta}{\sqrt{n}}\sum_{i=1}^{n} a_i - x_{n+1}\right)$ is normal inwards to X at x_{n+1} and so $\frac{a+\varepsilon}{\|a+\varepsilon\|} = -N(x_{n+1})$. By assumption on n and $\left\|\sum_{i=1}^{n}(a_i - a)\right\|$ we have

$$\|\varepsilon\| \leq \frac{1}{n}\left\|\sum_{i=1}^{n}(a_i - a)\right\| + \frac{D}{\eta\sqrt{n}} \leq \frac{\|a\|}{4} + \frac{D}{\eta\sqrt{16D^2/\eta^2\|a\|^2}} = \frac{\|a\|}{4} + \frac{\|a\|}{4} = \frac{\|a\|}{2}.$$

Hence Lemma P1 gives

$$\|N(x^*) - N(x_{n+1})\| \le 6\frac{\|\varepsilon\|}{\|a\|} \le \frac{6}{\|a\|}\left(\frac{1}{n}\Big\|\sum_{i=1}^{n}(a_i - a)\Big\| + \frac{\|x_{n+1}\|}{\eta\sqrt{n}}\right)$$
$$\le \frac{6}{\|a\|}\left(\frac{1}{n}\Big\|\sum_{i=1}^{n}(a_i - a)\Big\| + \frac{D}{\eta\sqrt{n}}\right)$$

Multiply both sides by $\|a\|/6$ to complete the proof. □

Lemma P3. Suppose $\big\|\sum_{i=1}^{N}(a_i - a)\big\| \le N\|a\|/2$. For $y^* = \operatorname{argmin}\big\{\sum_{i=1}^{N} a_i \cdot x : x \in X\big\}$ we have

$$\frac{\|a\|}{6}\|N(x^*) - N(y^*)\| \le \frac{1}{N}\Big\|\sum_{i=1}^{N}(a_i - a)\Big\|$$

Proof. Use Lemma P1 with $\varepsilon = \frac{1}{N}\sum_{i=1}^{N}(a_i - a)$ and how $-\frac{a}{\|a\|}$ and $-\frac{a+\varepsilon}{\|a+\varepsilon\|}$ are unit outwards normals to X at x^* and y^* respectively. □

Now we combine Proposition 2 with Lemma P3 to get tail bounds for Follow-the-Leader.

Lemma P4. For $y^* = \operatorname{argmin}\left\{\sum_{i=1}^{N} a_i \cdot x : x \in X\right\}$ and each $t \le \|a\|/2$ we have

$$P\left(\frac{\|a\|}{6}\|N(x^*) - N(y^*)\| > t\right) < 2\exp\left(-\frac{t^2}{2R^2}N\right).$$

Proof. For random variables $X_i = a_i - a$ Proposition 2 says $P\left(\big\|\frac{1}{N}\sum_{i=1}^{N}(a_i - a)\big\| \le t\right) \ge 1 - 2\exp\left(-\frac{t^2}{2R^2}N\right)$ for any $t > 0$. Moreover if the event $\big\{\big\|\frac{1}{N}\sum_{i=1}^{N}(a_i - a)\big\| \le t\big\}$ holds for some $t \le \frac{\|a\|}{2}$ we have $\frac{1}{N}\big\|\sum_{i=1}^{N}(a_i - a)\big\| \le \frac{\|a\|}{2}$ and so $\big\|\sum_{i=1}^{N}(a_i - a)\big\| \le N\frac{\|a\|}{2}$. Hence by Lemma P3 the event $\big\{\frac{\|a\|}{6}\|N(x^*) - N(y^*)\| \le t\big\}$ also holds. It follows $P\left(\frac{\|a\|}{6}\|N(x^*) - N(y^*)\| \le t\right) \ge P\left(\big\|\frac{1}{N}\sum_{i=1}^{N}(a_i - a)\big\| \le t\right) \ge 1 - 2\exp\left(-\frac{t^2}{2R^2}N\right)$. Take complements to complete the proof. □

Now we combine Proposition 2 with Lemma P2 to get tail bounds for Lazy Gradient Descent.

Lemma P5. Suppose $n \ge 16D^2/\eta^2\|a\|^2$. For the Online Lazy Gradient Descent actions $x_1, x_2, ...$ and each $t \le \|a\|/4$ we have

$$P\left(\frac{\|a\|}{6}\|N(x^*) - N(x_{n+1})\| > \frac{D}{\eta\sqrt{n}} + t\right) < 2\exp\left(-\frac{t^2}{2R^2}n\right).$$

Proof. For random variables $X_i = a_i - a$ Proposition 2 says $P\left(\big\|\frac{1}{n}\sum_{i=1}^{N}(a_i - a)\big\| \le t\right) \ge 1 - 2\exp\left(-\frac{t^2}{2R^2}n\right)$ for any $t > 0$. Moreover if the event $\big\{\big\|\frac{1}{n}\sum_{i=1}^{n}(a_i - a)\big\| \le t\big\}$ holds for some $t \le \frac{\|a\|}{4}$ we have $\frac{1}{n}\big\|\sum_{i=1}^{n}(a_i - a)\big\| \le \frac{\|a\|}{4}$ and so $\big\|\sum_{i=1}^{n}(a_i - a)\big\| \le n\frac{\|a\|}{2}$. Hence by Lemma P2 the event $\big\{\frac{\|a\|}{6}\|N(x^*) - N(x_{n+1})\| \le \frac{D}{\eta\sqrt{n}} + \frac{1}{n}\big\|\sum_{i=1}^{n}(a_i - a)\big\|.\big\}$ also holds. Since $\big\{\big\|\frac{1}{n}\sum_{i=1}^{n}(a_i - a)\big\| \le t\big\}$ holds so does $\big\{\frac{\|a\|}{6}\|N(x^*) - N(x_{n+1})\| \le \frac{D}{\eta\sqrt{n}} + t\big\}$. It follows $P\left(\frac{\|a\|}{6}\|N(x^*) - N(x_{n+1})\| \le \frac{D}{\eta\sqrt{n}} + t\right) \ge P\left(\big\|\frac{1}{n}\sum_{i=1}^{n}(a_i - a)\big\| \le t\right) \ge 1 - 2\exp\left(-\frac{t^2}{2R^2}n\right)$. Take complements to complete the proof. □

Lemmas P4 and P5 give tail bounds for $\|N(x^*)-N(y^*)\|$ and $\|N(x^*)-N(x_{n+1})\|$ respectively. For the main analysis we must use the lemmas to get tail bounds for $\|N(x^*)-N(y^*)\|^2$ and $\|N(x^*)-N(x_{n+1})\|^2$.

Lemma P6. Suppose the random variable Z takes values in $[0, W]$ and there exist $K, k, t_0 \geq 0$ such that for all $t \leq t_0$ we have $P(Z > K + t) < 2e^{-kt^2}$. Then we also have

$$\mathbb{E}[Z^2] \leq 2\left(K\sqrt{\frac{\pi}{k}} + \frac{1}{k}\right) + 2W^2 e^{-kt_0^2}$$

Proof. Let $\delta \geq 0$ be arbitrary and $t = \sqrt{\delta} - K$. Since Z is nonnegative we have $\{Z > K + t\} = \{Z^2 > (K + t)^2\} = \{Z^2 > \delta\}$. For $\delta \leq \delta_0 = (t_0 + K)^2$ we have $t \leq t_0$ and so $P(Z^2 > \delta) = P(Z > K + t) \leq 2e^{-kt^2} = 2e^{-k(\sqrt{\delta}-K)^2}$. Hence Lemma 11 gives

$$\begin{aligned}
\mathbb{E}[Z^2] &\leq \int_0^\infty P(Z^2 > \delta)d\delta = \int_0^{W^2} P(Z^2 > \delta)d\delta \\
&= \int_0^{(t_0+K)^2} P(Z^2 > \delta)d\delta + \int_{(t_0+K)^2}^{W^2} P(Z^2 > \delta)d\delta \\
&\leq 2\int_0^{(t_0+K)^2} \min\{1, 2e^{-k(\sqrt{\delta}-K)^2}\}d\delta + \int_{(t_0+K)^2}^{W^2} 2e^{-k(\sqrt{\delta_0}-K)^2}d\delta \\
&\leq \int_0^\infty \min\{1, 2e^{-k(\sqrt{\delta}-K)^2}\}d\delta + \int_{(t_0+K)^2}^{W^2} 2e^{-kt_0^2}d\delta \\
&\leq \int_0^\infty \min\{1, 2e^{-k(\sqrt{\delta}-K)^2}\}d\delta + 2W^2 e^{-kt_0^2}. \qquad (6)
\end{aligned}$$

For the integral write

$$\begin{aligned}
\int_0^\infty \min\{1, 2e^{-k(\sqrt{\delta}-K)^2}\}d\delta &= \int_0^{K^2} \min\{1, 2e^{-k(\sqrt{\delta}-K)^2}\}d\delta + \int_{K^2}^\infty \min\{1, 2e^{-k(\sqrt{\delta}-K)^2}\}d\delta \\
&\leq K^2 + 2\int_{K^2}^\infty e^{-k(\sqrt{\delta}-K)^2}d\delta.
\end{aligned}$$

To simplify the integral write $t = \sqrt{\delta}$. Then $dt = d\delta/2\sqrt{\delta}$ and $d\delta = 2\sqrt{\delta}dt = 2tdt$. Hence we get

$$2\int_{K^2}^\infty e^{-k(\sqrt{\delta}-K)^2}d\delta = 4\int_K^\infty te^{-k(t-K)^2}dt = 4\int_0^\infty (t + K)e^{-kt^2}dt = 2\left(K\sqrt{\frac{\pi}{k}} + \frac{1}{k}\right).$$

where we have used the Gaussian integral formula [24] and integration by parts to compute the integral. Combine with (6) to complete the proof. □

Now we combine Lemmas P5 and P6.

Lemma 7. Suppose $n \geq 16D^2/\eta^2\|a\|^2$. The Online Lazy Gradient Descent actions $x_1, x_2, ...$ give each

$$\mathbb{E}\|N(x^*)-N(x_{n+1})\|^2 \leq \frac{1}{n}\frac{36}{\|a\|^2}\left(\frac{\sqrt{8\pi}DR}{\eta} + 4R^2\right) + 8\exp\left(-\frac{\|a\|^2}{32R^2}n\right)$$

Proof. In the notation of Lemma P6 let

$$Z = \frac{\|a\|}{6}\|N(x^*)-N(x_{n+1})\| \qquad K = \frac{D}{\eta\sqrt{n}} \qquad k = \frac{n}{2R^2} \qquad t_0 = \frac{\|a\|}{4}.$$

Since $N(x^*), N(x_{n+1})$ are unit vectors $\max Z \leq \frac{\|a\|}{3} = W$. Combine Lemmas P5 and P6 and to bound $\mathbb{E}[Z^2]$. The the terms on the right-hand-side of Lemma P6 simplify to

$$\frac{2}{k} = \frac{4R^2}{n}$$
$$2K\sqrt{\frac{\pi}{k}} = \frac{2D}{\eta\sqrt{n}}\frac{\sqrt{2\pi}R}{\sqrt{n}} = \frac{1}{n}\frac{\sqrt{8\pi}DR}{\eta}$$
$$2W^2e^{-kt_0^2} = 2\frac{\|a\|^2}{9}\exp\left(-\frac{n}{2R^2}\frac{\|a\|^2}{16}\right) = \frac{2}{9}\|a\|^2\exp\left(-\frac{\|a\|^2}{32R^2}n\right).$$

To bound $\mathbb{E}\|N(x^*)-N(x_{n+1})\|^2$ multiply the bound for $\mathbb{E}[Z^2]$ by $\frac{36}{\|a\|^2}$. □

Similarly combine Lemmas P4 and P6 to get the following.

Lemma 11. For $y^* = \operatorname{argmin}\left\{\sum_{i=1}^N a_i \cdot x : x \in X\right\}$. The Online Lazy Gradient Descent actions give

$$\mathbb{E}\|N(x^*) - N(y^*)\|^2 \leq \frac{144R^2}{\|a\|^2 N} + 8\exp\left(-\frac{\|a\|^2}{8R^2}N\right).$$

Proof. In the notation of Lemma P6 let

$$Z = \frac{\|a\|}{6}\|N(y^*) - N(x^*)\| \qquad K = 0 \qquad k = \frac{n}{2R^2} \qquad t_0 = \frac{\|a\|}{2}.$$

Since $N(x^*), N(y^*)$ are unit vectors $\max Z \leq \frac{\|a\|}{3} = W$. Combine Lemmas P4 and P6 and to bound $\mathbb{E}[Z^2]$. The the terms on the right-hand-side of Lemma P6 simplify to

$$\frac{2}{k} = \frac{4R^2}{N} \qquad 2K\sqrt{\frac{\pi}{k}} = 0 \qquad 2W^2e^{-kt_0^2} = \frac{2}{9}\|a\|^2\exp\left(-\frac{\|a\|^2}{8R^2}N\right).$$

To bound $\mathbb{E}\|N(y^*)-N(x^*)\|^2$ multiply the bound for $\mathbb{E}[Z^2]$ by $\frac{36}{\|a\|^2}$. □

Lemma 8. For each neighborhood U of x^* in $\mathbb{R}^d$ the Online Lazy Gradient Descent actions $x_1, x_2, \ldots$ give $\sum_{i=1}^\infty P(x_i \notin U) < \infty$.

Proof. Lemma P5 says there are $K_1, k_1, \delta_1 > 0$ such that for all $\delta \leq \widetilde{\delta}_1$ we have

$$P\left(\|N(x_{n+1}) - N(x^*)\| > K_1/\sqrt{n} + \delta\right) \leq 2e^{-k_1 n\delta^2}$$

Proposition 1 and Lemma 2 combine to give

$$d(x_{n+1}, x^*) \leq \frac{2}{\|\nabla N(x^*)v\|}\|N(x_{n+1}) - N(x^*)\| \leq \frac{2}{m}\|N(x_{n+1}) - N(x^*)\|.$$

Hence there are $K_2, k_2, \delta_2 > 0$ such that for all $\delta \leq \delta_2$ we have

$$P\left(d(x_{n+1}, x^*) > K_2/\sqrt{n} + \delta\right) \leq 2e^{-k_1 n\delta^2} \tag{7}$$

Since U is a neighborhood it contains some ball $B(x^*, r)$ with respect to the geodesic distance. Thus it is enough to show the series $\sum_{i=1}^\infty P(d(x_{i+1} - x^*) > r)$ is finite. In particular it is enough to show some tail is finite. To that end let $n_0 \in \mathbb{N}$ be large enough that $K_2/\sqrt{n_0} \leq \frac{1}{2}\min\{r, \delta_2\}$. Plug $\delta = \frac{1}{2}\min\{r, \delta_2\}$ into (7) to get $K_2/\sqrt{n} + \delta \leq K_2/\sqrt{n_0} + \delta \leq \min\{r, \delta_2\} \leq r$. Hence we have

$$P\left(d(x_{n+1}, x^*) > r\right) \leq P\left(d(x_{n+1}, x^*) > K_2/\sqrt{n} + \delta\right) \leq 2e^{-k_1\delta^2 n}$$

Finally take the series. Since the terms are decreasing we can bound the series with the integral.

$$\sum_{i>n_0}^\infty P\left(d(x_{i+1}, x^*) > r\right) \leq 2\sum_{i>n_0}^\infty e^{-k_1\delta^2 i} \leq 2\int_{n_0}^\infty e^{-k_1\delta^2 x}dx \leq \frac{2}{k_2\delta^2}.$$

□