Techinical Proofs for
"Nonlinear Learning using Local Coordinate Coding"

1 Notations and Main Results

Definition 1.1 (Lipschitz Smoothness) *A function $f(x)$ on $\mathbb{R}^d$ is (α, β, p)-Lipschitz smooth with respect to a norm $\|\cdot\|$ if*

$$|f(x') - f(x)| \le \alpha \|x - x'\|,$$

and

$$|f(x') - f(x) - \nabla f(x)^\top (x' - x)| \le \beta \|x - x'\|^{1+p},$$

where we assume $\alpha, \beta > 0$ and $p \in (0, 1]$.

Definition 1.2 (Coordinate Coding) *A coordinate coding is a pair (γ, C), where $C \subset \mathbb{R}^d$ is a set of anchor points, and γ is a map of $x \in \mathbb{R}^d$ to $[\gamma_v(x)]_{v \in C} \in R^{|C|}$ such that $\sum_v \gamma_v(x) = 1$. It induces the following physical approximation of x in $\mathbb{R}^d$:*

$$\gamma(x) = \sum_{v \in C} \gamma_v(x) v.$$

Moreover, for all $x \in \mathbb{R}^d$, we define the coding norm as

$$\|x\|_\gamma = \left(\sum_{v \in C} \gamma_v(x)^2 \right)^{1/2}.$$

Proposition 1.1 *The map $x \to \sum_{v \in C} \gamma_v(x) v$ is invariant under any shift of the origin for representing data points in $\mathbb{R}^d$ if and only if $\sum_v \gamma_v(x) = 1$.*

Lemma 1.1 (Linearization) *Let (γ, C) be an arbitrary coordinate coding on $\mathbb{R}^d$. Let f be an (α, β, p)-Lipschitz smooth function. We have for all $x \in \mathbb{R}^d$:*

$$\left| f(x) - \sum_{v \in C} \gamma_v(x) f(v) \right| \le \alpha \|x - \gamma(x)\| + \beta \sum_{v \in C} |\gamma_v(x)| \, \|v - \gamma(x)\|^{1+p}.$$

Definition 1.3 (Localization Measure) *Given α, β, p, and coding (γ, C), we define*

$$Q_{\alpha,\beta,p}(\gamma, C) = \mathbb{E}_x \left[\alpha \|x - \gamma(x)\| + \beta \sum_{v \in C} |\gamma_v(x)| \, \|v - \gamma(x)\|^{1+p} \right].$$

Definition 1.4 (Manifold) *A subset $\mathcal{M} \subset \mathbb{R}^d$ is called a p-smooth ($p > 0$) manifold with intrinsic dimensionality $m = m(\mathcal{M})$ if there exists a constant $c_p(\mathcal{M})$ such that given any $x \in \mathcal{M}$, there exists m vectors $v_1(x), \ldots, v_m(x) \in \mathbb{R}^d$ so that $\forall x' \in \mathcal{M}$:*

$$\inf_{\gamma \in \mathbb{R}^m} \left\| x' - x - \sum_{j=1}^{m} \gamma_j v_j(x) \right\| \leq c_p(\mathcal{M}) \|x' - x\|^{1+p}.$$

Definition 1.5 (Covering Number) *Given any subset $\mathcal{M} \subset \mathbb{R}^d$, and $\epsilon > 0$. The covering number, denoted as $\mathcal{N}(\epsilon, \mathcal{M})$, is the smallest cardinality of an ϵ-cover $C \subset \mathcal{M}$. That is,*

$$\sup_{x \in \mathcal{M}} \inf_{v \in C} \|x - v\| \leq \epsilon.$$

Theorem 1.1 (Manifold Coding) *If the data points x lie on a compact p-smooth manifold $\mathcal{M}$, and the norm is defined as $\|x\| = (x^\top A x)^{1/2}$ for some positive definite matrix A. Then given any $\epsilon > 0$, there exist anchor points $C \subset \mathcal{M}$ and coding γ such that*

$$\begin{aligned} &|C| \leq (1 + m(\mathcal{M}))\mathcal{N}(\epsilon, \mathcal{M}), \\ &Q_{\alpha,\beta,p}(\gamma, C) \leq [\alpha c_p(\mathcal{M}) + (1 + \sqrt{m} + 2^{1+p}\sqrt{m})\beta]\, \epsilon^{1+p}. \end{aligned}$$

Moreover, for all $x \in \mathcal{M}$, we have $\|x\|_\gamma^2 \leq 1 + (1 + \sqrt{m})^2$.

Given a local-coordinate coding scheme (γ, C), we approximate each $f(x) \in \mathcal{F}^a_{\alpha,\beta,p}$ by

$$f(x) \approx f_{\gamma,C}(\hat{w}, x) = \sum_{v \in C} \hat{w}_v \gamma_v(x),$$

where we estimate the coefficients using ridge regression as:

$$[\hat{w}_v] = \arg\min_{[w_v]} \left[\sum_{i=1}^{n} \phi\left(f_{\gamma,C}(w, x_i), y_i\right) + \lambda \sum_{v \in C} (w_v - g(v))^2 \right], \tag{1}$$

Theorem 1.2 (Generalization Bound) *Suppose $\phi(p, y)$ is Lipschitz: $|\phi'_1(p, y)| \leq B$. Consider coordinate coding (γ, C), and the estimation method (1) with random training examples $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$. Then the expected generalization error satisfies the inequality:*

$$\begin{aligned} &\mathbb{E}_{S_n}\, \mathbb{E}_{x,y} \phi(f_{\gamma,C}(\hat{w}, x), y) \\ \leq &\inf_{f \in \mathcal{F}_{\alpha,\beta,p}} \left[\mathbb{E}_{x,y} \phi\left(f(x), y\right) + \lambda \sum_{v \in C} (f(v) - g(v))^2 \right] + \frac{B^2}{2\lambda n} \mathbb{E}_x \|x\|_\gamma^2 + B Q_{\alpha,\beta,p}(\gamma, C). \end{aligned}$$

Theorem 1.3 (Consistency) *Suppose the data lie on a compact manifold $\mathcal{M} \subset \mathbb{R}^d$, and the norm $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$. If loss function $\phi(p, y)$ is Lipschitz. As $n \to \infty$, we choose $\alpha, \beta \to \infty$, $\alpha/n, \beta/n \to 0$ (α, β depends on n), and $p = 0$. Then it is possible to find coding (γ, C) using unlabeled data such that $|C|/n \to 0$ and $Q_{\alpha,\beta,p}(\gamma, C) \to 0$. If we pick $\lambda n \to \infty$, and $\lambda |C| \to 0$. Then the local coordinate coding method (1) with $g(v) \equiv 0$ is consistent as $n \to \infty$:*

$$\lim_{n \to \infty} \mathbb{E}_{S_n}\, \mathbb{E}_{x,y} \phi(f(\hat{w}, x), y) = \inf_{f : \mathcal{M} \to \mathbb{R}} \mathbb{E}_{x,y} \phi\left(f(x), y\right).$$

2 Proofs

2.1 Proof of Proposition 1.1

Consider a change of the $\mathbb{R}^d$ origin by $u \in \mathbb{R}^d$, which shifts any point $x \in \mathbb{R}^d$ to $x+u$, and points $v \in C$ to $v+u$. The shift-invariance requirement implies that after the change, we map $x+u$ to $\sum_{v\in C}\gamma_v(x)v + u$, which should equal $\sum_{v\in C}\gamma_v(x)(v+u)$. This is equivalent to $u = \sum_{v\in C}\gamma_v(x)u$, which holds if and only if $\sum_{v\in C}\gamma_v(x) = 1$.

2.2 Proof of Lemma 1.1

For simplicity, let $\gamma_v = \gamma_v(x)$ and $x' = \gamma(x) = \sum_{v\in C}\gamma_v v$. We have

$$\begin{aligned}
|f(x) - \sum_{v\in C}\gamma_v f(v)| \leq & |f(x) - f(x')| + \left|\sum_{v\in C}\gamma_v(f(v) - f(x'))\right| \\
= & |f(x) - f(x')| + \left|\sum_{v\in C}\gamma_v(f(v) - f(x') - \nabla f(x')^\top (v - x'))\right| \\
\leq & |f(x) - f(x')| + \sum_{v\in C}|\gamma_v||(f(v) - f(x') - \nabla f(x')^\top (v - x'))| \\
\leq & \alpha\|x - x'\|_2 + \beta\sum_{v\in C}|\gamma_v|\|x' - v\|^{1+p}.
\end{aligned}$$

This implies the bound.

2.3 Proof of Theorem 1.1

Let $m = m(\mathcal{M})$. Given any $\epsilon > 0$, consider an ϵ-cover C' of $\mathcal{M}$ with $|C'| \leq \mathcal{N}(\epsilon, \mathcal{M})$. Given each $u \in C'$, define $C_u = \{v_1(u), \ldots, v_d(u)\}$, where $v_j(u)$ are defined in Definition 1.4. Define the anchor points as

$$C = \cup_{u\in C'}\{u + v_j(u) : j = 1, \ldots, m\} \cup C'.$$

It follows that $|C| \leq (1+m)\mathcal{N}(\epsilon, \mathcal{M})$.

In the following, we only need to prove the existence of a coding γ on $\mathcal{M}$ that satisfies the requirement of the theorem. Without loss of generality, we assume that $\|v_j(u)\| = \epsilon$ for each u and j, and given u, $\{v_j(u) : j = 1, \ldots, m\}$ are orthogonal with respect to A: $v_j^\top(u)Av_k(u) = 0$ when $j \neq k$.

For each $x \in \mathcal{M}$, let $u_x \in C'$ be the closest point to x in C'. We have $\|x - u_x\| \leq \epsilon$ by the definition of C'. Now, Definition 1.4 implies that there exists $\gamma'_j(x)$ ($j = 1, \ldots, m$) such that

$$\left\|x - u_x - \sum_{j=1}^{m}\gamma'_j(x)v_j(u_x)\right\| \leq c_p(\mathcal{M})\epsilon^{1+p}.$$

The optimal choice is the A-projection of $x - u_x$ to the subspace spanned by $\{v_j(u_x) : j = 1, \ldots, m\}$. The orthogonality condition thus implies that

$$\sum_{j=1}^{m}\gamma'_j(x)^2\|v_j(u_x)\|^2 \leq \|x - u_x\|^2 \leq \epsilon^2.$$

Therefore

$$\sum_{j=1}^{m} \gamma_j'(x)^2 \leq 1,$$

which implies that for all x:

$$\sum_{j=1}^{m} |\gamma_j'(x)| \leq \sqrt{m}.$$

We can now define the coordinate coding of $x \in \mathcal{M}$ as

$$\gamma_v(x) = \begin{cases} \gamma_j' & v = u_x + v_j(u_x) \\ 1 - \sum_{j=1}^{m} \gamma_j' & v = u_x \\ 0 & \text{otherwise} \end{cases} .$$

This implies the following bounds:

$$\|x - \gamma(x)\| \leq c_p(\mathcal{M})\epsilon^{1+p}$$

and

$$\sum_{v \in C} |1 - \sum j = 1^m \gamma_j'| \, \|v - \gamma(x)\|^{1+p} = |\gamma_{u_x}(x)| \|\gamma(x) - u_x\| + \sum_{j=1}^{m} |\gamma_j'(x)| \, \|(v - u_x) - (\gamma(x) - u_x)\|^{1+p} \tag{2}$$

$$\leq (1 + \sqrt{m})\epsilon^{1+p} \sum_{j=1}^{m} |\gamma_j'(x)|(\epsilon + \epsilon)^{1+p} \tag{3}$$

$$= [1 + \sqrt{m} + 2^{1+p}\sqrt{m}]\epsilon^{1+p}. \tag{4}$$

where we have used $\|v - u_x\| = \epsilon$, and $\|\gamma(x) - u_x\| \leq \|x - u_x\| \leq \epsilon$.

2.4 Proof of Theorem 1.2

Consider $n+1$ samples $S_{n+1} = \{(x_1, y_1), \ldots, (x_{n+1}, y_{n+1})\}$. We shall introduce the following notation:

$$[\tilde{w}_v] = \arg\min_{[w_v]} \left[\frac{1}{n} \sum_{i=1}^{n+1} \phi\left(f_{\gamma,C}(w, x_i), y_i\right) + \lambda \sum_{v \in C} w_v^2 \right]. \tag{5}$$

Let k be an integer randomly drawn from $\{1, \ldots, n+1\}$. Let $[\hat{w}_v^{(k)}]$ be the solution of

$$[\hat{w}_v^{(k)}] = \arg\min_{[w_v]} \left[\frac{1}{n} \sum_{i=1,\ldots,n+1; i \neq k} \phi\left(f_{\gamma,C}(w, x_i), y_i\right) + \lambda \sum_{v \in C} w_v^2 \right],$$

with the k-th example left-out.

We have the following stability lemma from [1], which can be stated as follows using our terminology:

Lemma 2.1 *The following inequality holds*

$$|f_{\gamma,C}(\hat{w}^{(k)}, x_k) - f_{\gamma,C}(\tilde{w}, x_k)| \leq \frac{\|x_k\|_\gamma^2}{2\lambda n} |\phi_1'(f_{\gamma,C}(\tilde{w}, x_k), y_k)|.$$

By using Lemma 2.1, we obtain for all $\alpha > 0$:

$$\begin{aligned}
&\phi(f_{\gamma,C}(\tilde{w}, x_k), y_k) - \phi(f_{\gamma,C}(\hat{w}^{(k)}, x_k), y_k) \\
=&\phi(f_{\gamma,C}(\tilde{w}, x_k), y_k) - \phi(f_{\gamma,C}(\hat{w}^{(k)}, x_k), y_k) - \phi_1'(f_{\gamma,C}(\hat{w}^{(k)}, x_k), y_k)(f_{\gamma,C}(\tilde{w}, x_k) - f_{\gamma,C}(\hat{w}^{(k)}, x_k)) \\
&\quad + \phi_1'(f_{\gamma,C}(\hat{w}^{(k)}, x_k), y_k)(f_{\gamma,C}(\tilde{w}, x_k) - f_{\gamma,C}(\hat{w}^{(k)}, x_k)) \\
\geq&\phi_1'(f_{\gamma,C}(\hat{w}^{(k)}, x_k), y_k)(f_{\gamma,C}(\tilde{w}, x_k) - f_{\gamma,C}(\hat{w}^{(k)}, x_k)) \\
\geq& - \phi_1'(f_{\gamma,C}(\hat{w}^{(k)}, x_k), y_k)^2 \|x_k\|_\gamma^2 / (2\lambda n) \\
\geq& - B^2 \|x_k\|_\gamma^2 / (2\lambda n).
\end{aligned}$$

In the above derivation, the first inequality uses the convexity of $\phi(f, y)$ with respect to f, which implies that $\phi(f_1, y) - \phi(f_2, y) - \phi_1'(f_2, y)(f_1 - f_2) \geq 0$. The second inequality uses Lemma 2.1, and the third inequality uses the assumption of the loss function.

Now by summing over k, and consider any fixed $f \in \mathcal{F}_{\alpha,\beta,p}$, we obtain:

$$\begin{aligned}
&\sum_{k=1}^{n+1} \phi(f_{\gamma,C}(\hat{w}^{(k)}, x_k), y_k) \\
\leq&\sum_{k=1}^{n+1} \left[\phi(f_{\gamma,C}(\tilde{w}, x_k), y_k) + \frac{B}{2\lambda n} \|x_k\|_\gamma^2 \right] \\
\leq& n \left[\frac{1}{n} \sum_{k=1}^{n+1} \phi\left(\sum_{v \in C} \gamma_v(x_k) f(v), y_k \right) + \lambda \sum_{v \in C} f(v)^2 \right] + \frac{B^2}{2\lambda n} \sum_{k=1}^{n+1} \|x_k\|_\gamma^2 \\
\leq& n \left[\frac{1}{n} \sum_{k=1}^{n+1} [\phi\left(f(x_k), y_k\right) + BQ(x_k)] + \lambda \sum_{v \in C} f(v)^2 \right] + \frac{B^2}{2\lambda n} \sum_{k=1}^{n+1} \|x_k\|_\gamma^2,
\end{aligned}$$

where $Q(x) = \alpha \|x - \gamma(x)\| + \beta \sum_{v \in C} |\gamma_v(x)| \, \|v - \gamma(x)\|^{1+p}$. In the above derivation, the second inequality follows from the definition of $\tilde{w}$ as the minimizer of (5). The third inequality follows from Lemma 1.1. Now by taking expectation with respect to S_{n+1}, we obtain

$$\begin{aligned}
&(n+1)\mathbb{E}_{S_{n+1}} \phi(f_{\gamma,C}(\hat{w}^{(n+1)}, x_{n+1}), y_{n+1}) \\
\leq& n \left[\frac{n+1}{n} \mathbb{E}_{x,y} \phi\left(f(x), y\right) + \frac{n+1}{n} BQ_{\alpha,\beta,p}(\gamma, C) + \lambda \sum_{v \in C} f(v)^2 \right] + \frac{B^2(n+1)}{2\lambda n} \mathbb{E}_x \|x\|_\gamma^2.
\end{aligned}$$

This implies the desired bound.

2.5 Proof of Theorem 1.3

Note that any measurable function $f : \mathcal{M} \to R$ can be approximated by $\mathcal{F}_{\alpha,\beta,p}$ with $\alpha, \beta \to \infty$ and $p = 0$. Therefore we only need to show

$$\lim_{n \to \infty} \mathbb{E}_{S_n} \mathbb{E}_{x,y} \phi(f_{\gamma,C}(\hat{w}, x), y) = \lim_{n \to \infty} \inf_{f \in \mathcal{F}_{\alpha,\beta,p}} \mathbb{E}_{x,y} \phi\left(f(x), y\right).$$

Theorem 1.1 implies that it is possible to pick (γ, C) such that $|C|/n \to 0$ and $Q_{\alpha,\beta,p}(\gamma, C) \to 0$. Moreover, $\|x\|_\gamma$ is bounded.

Given any $f \in \mathcal{F}_{\alpha,\beta,0}$ and any n independent fixed $A > 0$; if we let $f_A(x) = \max(\min(f(x), A), -A)$, then it is clear that $f_A(x) \in \mathcal{F}_{\alpha,\alpha+\beta,0}$. Therefore Theorem 1.2 implies that as $n \to \infty$,

$$\mathbb{E}_{S_n}\, \mathbb{E}_{x,y}\phi(f_{\gamma,C}(\hat{w}, x), y) \leq \mathbb{E}_{x,y}\phi\left(f_A(x), y\right) + o(1).$$

Since A is arbitrary, we let $A \to \infty$ to obtain the desired result.

References

[1] Tong Zhang. Leave-one-out bounds for kernel methods. *Neural Computation*, 15:1397 – 1437, 2003.