Tractable Landscapes for Nonconvex Optimization

Shihua Zhang

November 3, 2021

Overview

Background

- Deep learning relies on optimizing a nonconvex loss.
- Even simple algorithms such as gradient descent often optimize the objective value to zero or near-zero.
- **Goal**: How to optimizate the nonconvex landscapes efficiently and identify their properties (for machine learning models)?
- Only apply to simpler nonconvex problems than deep learning.
- How to analyze deep learning with such landscape analysis is still open.

Global and Local Minimum

Definition 1 (Global/Local minimum)

1. For an objective function $f(w) : \mathbb{R}^d \to \mathbb{R}$, a point w^* is a global minimum if for every w we have $f(w^*) \leqslant f(w)$.
2. A point w is a local minimum/maximum if there exists a radius $\epsilon > 0$ such that for every $\|w' - w\|_2 \leqslant \epsilon$, we have $f(w) \leqslant f(w')$ ($f(w) \geqslant f(w')$ for local maximum).
3. A point w with $\nabla f(w) = 0$ is called a critical point, and for smooth functions all local minimum/maximum are critical points.

- Here we work with functions whose global minimum exists, and use $f(w^*)$ to denote its optimal value.

Spurious Local Minimum

Definition 2 (Spurious local minimum)

For an objective function $f(w) : \mathbb{R}^d \to \mathbb{R}$, a point w is a spurious local minimum if it is a local minimum, but $f(w) > f(w^*)$.

- Many optimization algorithms are based on the idea of local search, thus cannot escape from a spurious local minimum.
- Many noncovex objectives do not have spurious local minima.

Saddle Points

Definition 3 (Saddle point)

For an objective function $f(w) : \mathbb{R}^d \to \mathbb{R}$, a point w is a saddle point if $\nabla f(w) = 0$, and for every radius $\epsilon > 0$, there exists w^+, w^- within distance ϵ of w such that $f\left(w^-\right) < f(w) < f\left(w^+\right)$.

- This definition covers all cases but makes it very hard to verify whether a point is a saddle point.
- In most cases, it is possible to tell whether a point is a saddle point, local minimum or local maximum based on its Hessian.

Second Order Sufficient Condition

Theorem 4

For an objective function $f(w) : \mathbb{R}^d \to \mathbb{R}$ *and a critical point* w ($\nabla f(w) = 0$)*, we know:*
- If $\nabla^2 f(w) \succ 0$*,* w *is a local minimum.*
- If $\nabla^2 f(w) \prec 0$*,* w *is a local maximum.*
- If $\nabla^2 f(w)$ *has both a positive and a negative eigenvalue,* w *is a saddle point.*

- **Proof Hint**: looking at the second-order Taylor expansion.
- The three cases do not cover all the possible Hessian matrices.

Flat Regions

- **Challenge**: Even if a function does not have any spurious local minima or saddle point, it can still be hard to optimize.
- **Difficulty**: even if the norm $\|\nabla f(w)\|_2$ is small, unlike convex functions, one cannot conclude that $f(w)$ is close to $f(w^*)$.

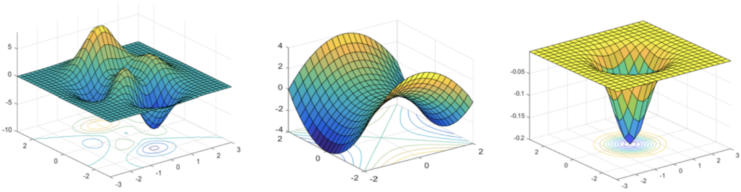

Figure: Obstacles for nonconvex optimization. From left to right: local minimum, saddle point and flat region.

Overview

Cases With a Unique Global Minimum

We first consider the case that is similar to convex objectives.

- The objective functions we look at have no spurious local minima or saddle points.
- **Obstacle**: points with small gradients may not be near-optimal.
- **Main Idea**: identify properties of the objective function, such that it keeps decreasing during the optimization process.

Cases With a Unique Global Minimum

Definition 5

Let $f(w)$ be an objective function with a unique global minimum w^*, then:

Polyak-Lojasiewicz: f satisfies Polyak-Lojasiewicz if there exists a value $\mu > 0$ such that for every w, $\|\nabla f(w)\|_2^2 \geqslant \mu\left(f(w) - f\left(w^*\right)\right)$

Weakly-quasi-convex: f is weakly-quasi-convex if there exists a value $\mu > 0$ such that for every w, $\langle \nabla f(w), w - w^* \rangle \geqslant \mu\left(f(w) - f\left(w^*\right)\right)$

Restricted Secant Inequality (RSI): f satisfies RSI if there exists a value μ such that for every w, $\langle \nabla f(w), w - w^* \rangle \geqslant \mu \|w - w^*\|_2^2$

Cases With a Unique Global Minimum

Any one of these three properties can imply fast convergence together with some smoothness of f.

Theorem 6

If an objective function f satisfies one of Polyak-Lojasiewicz, weakly-quasi-convex or RSI, and f is smooth, then gradient descent converges to global minimum with a geometric rate.

- Polyak-Lojasiewicz and RSI requires standard smoothness, weakly-quasi-convex requires a special smoothness property detailed in [Hardt et al., 2016].
- We will use generalized linear model (GLM) as an example to show how some of these properties can be used.

Generalized Linear Model

In GLM ([Kalai and Sastry, 2009], [Kakade et al., 2011]), the input consists of samples $\left\{x^{(i)}, y^{(i)}\right\}$ that are drawn from a distribution $\mathcal{D}$, where $(x, y) \sim \mathcal{D}$ satisfies

$$y = \sigma\left(w_*{}^\top x\right) + \epsilon$$

Generalized Linear Model

In GLM ([Kalai and Sastry, 2009], [Kakade et al., 2011]), the input consists of samples $\left\{x^{(i)}, y^{(i)}\right\}$ that are drawn from a distribution $\mathcal{D}$, where $(x, y) \sim \mathcal{D}$ satisfies

$$y = \sigma\left({w_*}^\top x\right) + \epsilon$$

- $\sigma : \mathbb{R} \to \mathbb{R}$ is a known monotone function, ϵ is a noise that satisfies $\mathbb{E}[\epsilon \mid x] = 0$.
- Consider Expected loss: $L(w) = \frac{1}{2} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\left(y - \sigma\left(w^\top x\right)^2\right)\right]$.

Generalized Linear Model

In GLM ([Kalai and Sastry, 2009], [Kakade et al., 2011]), the input consists of samples $\left\{x^{(i)}, y^{(i)}\right\}$ that are drawn from a distribution $\mathcal{D}$, where $(x, y) \sim \mathcal{D}$ satisfies

$$y = \sigma\left({w_*}^\top x\right) + \epsilon$$

- $\sigma : \mathbb{R} \to \mathbb{R}$ is a known monotone function, ϵ is a noise that satisfies $\mathbb{E}[\epsilon \mid x] = 0$.
- Consider Expected loss: $L(w) = \frac{1}{2} \underset{(x,y)\sim\mathcal{D}}{\mathbb{E}}\left[\left(y - \sigma\left(w^\top x\right)^2\right)\right]$.
- GLM: learn a single neuron where σ is its nonlinearity.

Generalized Linear Model

How to prove prop. (e.g., weakly-quasi-convex or RSI) for GLM?

The objective is rewritten as:

$$\begin{aligned} L(w) &= \frac{1}{2} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\left(y - \sigma\left(w^\top x\right)^2 \right) \right] \\ &= \frac{1}{2} \mathop{\mathbb{E}}_{(x,\epsilon)} \left[\left(\epsilon + \sigma\left(w_*^\top x\right) - \sigma\left(w^\top x\right) \right)^2 \right] \\ &= \frac{1}{2} \mathop{\mathbb{E}}_{\epsilon} \left[\epsilon^2\right] + \frac{1}{2} \mathop{\mathbb{E}}_{x} \left[\left(\sigma\left(w_*^\top x\right) - \sigma\left(w^\top x\right)^2 \right) \right]. \end{aligned} \tag{1}$$

- This decomposition is helpful as $\frac{1}{2}\mathop{\mathbb{E}}_{\epsilon}\left[\epsilon^2\right]$ is just a constant.

Generalized Linear Model

Consider the derivative of the objective:

$$\nabla L(w) = \mathbb{E}_x \left[\left(\sigma\left(w^\top x\right) - \sigma\left(w_*^\top x\right) \right) \sigma'\left(w^\top x\right) x \right]. \tag{2}$$

Then we have:

$$\begin{aligned}
&\langle \nabla L(w), w - w^* \rangle \\
=&\mathbb{E}_x \left[\left(\sigma\left(w^\top x\right) - \sigma\left(w_*^\top x\right) \right) \sigma'\left(w^\top x\right) \left(w^\top x - w_*^\top x\right) \right] \\
=&\mathbb{E}_x \left[\sigma'(\xi) \sigma'\left(w^\top x\right) \left(w^\top x - w_*^\top x\right)^2 \right].
\end{aligned} \tag{3}$$

By making more assumptions on σ and the distribution of x, it is possible to lowerbound $\langle \nabla L(w), w - w^* \rangle$ in the form required by either weakly-quasi-convex or RSI.

Overview

Permutation Symmetry for Neural Networks

Consider a two-layer neural network $h_\theta(x) : \mathbb{R}^d \to \mathbb{R}$. The parameters θ is $(w_1, w_2, \cdots, w_k)$.

- The function can be evaluated as $h_\theta(x) = \sum_{i=1}^{k} \sigma(\langle w_i, x \rangle)$.
- Given a dataset $\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right) \overset{\text{i.i.d.}}{\sim} \mathcal{D}$.
- The objective $f(\theta) = L(h_\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell((x, y), h_\theta)]$ has permutation symmetry.
- That is, for any permutation $\pi(\theta)$ that permutes the weights of the neurons, $f(\theta) = f(\pi(\theta))$.

Permutation Symmetry for Neural Networks

The permutation symmetry has many implications:

- If the global minimum θ^* is a point where not all neurons have the same weight, then there must be equivalent global minimum $f(\pi(\theta^*))$ for every permutation π.
- An objective with this symmetry must also be nonconvex, because if it were convex, the point $\bar{\theta} = \frac{1}{k!}\sum_{\pi}\pi(\theta^*)$ must be a global minimum.
- However, for $\bar{\theta}$ the weight vectors of the neurons are all equal to $\frac{1}{k}\sum_{i=1}^{k} w_i$, so $h_{\bar{\theta}}(x) = k\sigma\left(\left\langle \frac{1}{k}\sum_{i=1}^{k} w_i, x\right\rangle\right)$ is equivalent to a neural network with a single neuron.

Permutation Symmetry for Neural Networks

The permutation symmetry has many implications:

- f must be nonconvex.
- It is also possible to show that functions with symmetry must have saddle points.

- To optimize f, the algorithm needs to be able to either avoid or escape from saddle points.
- More concretely, one would like to find a second order stationary point.

Second order stationary point (SOSP)

Definition 7 (Second order stationary point (SOSP))

For an objective function $f(w) : \mathbb{R}^d \to \mathbb{R}$, a point w is a second order stationary point if $\nabla f(w) = 0$ and $\nabla^2 f(w) \succeq 0$

- The conditions for SOSP are known as the second order necessary conditions for a local minimum.
- The optimization algorithms can be used to find an approximate second order stationary point.

Definition 8 (Approximate second order stationary point)

For an objective function $f(w) : \mathbb{R}^d \to \mathbb{R}$, a point w is a (ϵ, γ)-second order stationary point ((ϵ, γ)-SOSP) if $\|\nabla f(w)\|_2 \leqslant \epsilon$ and $\lambda_{\min}\left(\nabla^2 f(w)\right) \geqslant -\gamma$

Locally Optimizable

Define a class of functions that can be optimized efficiently and allow symmetry and saddle points.

Definition 9 (Locally optimizable functions)

An objective function $f(w)$ is locally optimizable, if for every $\tau > 0$, there exists $\epsilon, \gamma = \text{poly}(\tau)$ such that every $(\epsilon, \gamma) - \text{SOSP}$ w of f satisfies $f(w) \leqslant f(w_*) + \tau$.

- Roughly speaking, an objective function is locally optimizable if every local minimum of the function is also a global minimum, and the Hessian of every saddle point has a negative eigenvalue.

Locally Optimizable Functions

Locally optimizable objective functions:

- Matrix sensing [Hardt et al., 2016]
- Matrix completion [Ge et al., 2016]
- Dictionary learning [Sun et al., 2016]
- Tensor decomposition [Ge et al., 2015]
- Certain objective for two-layer neural network [Ge et al., 2017]

Overview

Top Eigenvector of a Matrix

Here we look at a simple example of a locally optimizable function.

- Given a symmetric PSD matrix $M \in \mathbb{R}^{d \times d}$, the goal is to find its top eigenvector.
- More precisely, using SVD we can write M as

$$M = \sum_{i=1}^{d} \lambda_i v_i v_i^\top$$

  Here v_i's are orthonormal vectors that are eigenvectors of M, and λ_i's are the eigenvalues.
- For simplicity, we assume $\lambda_1 > \lambda_2 \geqslant \lambda_3 \geqslant \cdots \geqslant \lambda_d > 0$

Top Eigenvector of a Matrix

There are many objective functions whose global optima give the top eigenvector.

- For PSD matrix M, the global optima of

$$\max_{\|x\|_2=1} x^\top M x$$

  is the top eigenvector of M. However, this formulation requires a constraint.

- We instead work with an unconstrained version whose correctness follows from Eckart-Young Theorem:

$$\min_{x\in\mathbb{R}^d} f(x) := \frac{1}{4}\left\|M - xx^\top\right\|_F^2$$

Top Eigenvector of a Matrix

Consider the following unconstrained optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{4} \left\| M - xx^\top \right\|_F^2$$

- This function does have a symmetry in the sense that $f(x) = f(-x)$.
- Under the assumptions, the only global minima of it are $x = \pm\sqrt{\lambda_1} v_1$. They are the only 2^{nd}-order stationary points.
- Two proof strategies:
  - Characterizing all critical points
  - Finding directions of improvements

Characterizing All Critical Points

The first idea is simple:

- Solve the Eq. $\nabla f(x) = 0$ to get the position of all critical points.

Characterizing All Critical Points

The first idea is simple:

- Solve the Eq. $\nabla f(x) = 0$ to get the position of all critical points.

- For the critical points that are not the desired global minimum, try to prove that they are local maximum or saddle points.

Computing Gradient and Hessian

Expand $f(x+\delta)$ as follows:

$$
\begin{aligned}
f(x+\delta) =& \frac{1}{4}\left\|M-(x+\delta)(x+\delta)^{\top}\right\|_F^2 \\
=& \frac{1}{4}\left\|M-xx^{\top}-\left(x\delta^{\top}+\delta x^{\top}\right)-\delta\delta^{\top}\right\|_F^2 \\
=& \frac{1}{4}\left\|M-xx^{\top}\right\|_F^2-\frac{1}{2}\left\langle M-xx^{\top}, x\delta+\delta x^{\top}\right\rangle \\
&+\left[\frac{1}{4}\left\|x\delta^{\top}+\delta x^{\top}\right\|_F^2-\frac{1}{2}\left\langle M-xx^{\top}, \delta\delta^{\top}\right\rangle\right]+o\left(\|\delta\|_2^2\right)
\end{aligned}
$$

Thus we have:

$$
\nabla f(x)=\left(xx^{\top}-M\right)x, \quad \nabla^2 f(x)=\|x\|_2^2 I+2xx^{\top}-M
$$

Characterizing Critical Points

Set $\nabla f(x) = 0$, we have:

$$Mx = xx^\top x = \|x\|_2^2 x$$

- The only solutions to $Mx = \lambda x$ are if λ is an eigenvalue and x is (a scaled version) of the corresponding eigenvector.
- $x = \pm\sqrt{\lambda_i} v_i$ or $x = 0$. And $x = \pm\sqrt{\lambda_1} v_1$ are intended solutions.

Characterizing Critical Points

Set $\nabla f(x) = 0$, we have:

$$Mx = xx^\top x = \|x\|_2^2 x$$

- The only solutions to $Mx = \lambda x$ are if λ is an eigenvalue and x is (a scaled version) of the corresponding eigenvector.
- $x = \pm\sqrt{\lambda_i} v_i$ or $x = 0$. And $x = \pm\sqrt{\lambda_1} v_1$ are intended solutions.
- Next we need to show for every other critical point, its Hessian has a negative direction, i.e., there exists a δ such that $\delta^\top [\nabla^2 f(x)] \delta < 0$.
- Key: The main step of the proof involves guessing what is this direction δ. In this case we will choose $\delta = v_1$.

Characterizing Critical Points

When $x = \pm\sqrt{\lambda_i} v_i$, and $\delta = v_1$, we have:

$$\delta^\top \left[\nabla^2 f(x)\right] \delta = v_1^\top \left[\left\| \sqrt{\lambda_i} v_i \right\|_2^2 I + 2\lambda_i v_i v_i^\top - M \right] v_1 = \lambda_i - \lambda_1 < 0$$

The proof for $x = 0$ is very similar.
Combining all the steps above, we proved the theorem:

Theorem 10 (Properties of critical points)

The only critical points of $f(x)$ are of the form $x = \pm\sqrt{\lambda_i} v_i$ or $x = 0$. For all critical points except $x = \pm\sqrt{\lambda_1} v_1$, $\nabla^2 f(x)$ has a negative eigenvalue.

Characterizing Critical Points

When $x = \pm\sqrt{\lambda_i} v_i$, and $\delta = v_1$, we have:

$$\delta^\top \left[\nabla^2 f(x)\right] \delta = v_1^\top \left[\left\|\sqrt{\lambda_i} v_i\right\|_2^2 I + 2\lambda_i v_i v_i^\top - M\right] v_1 = \lambda_i - \lambda_1 < 0$$

The proof for $x = 0$ is very similar.
Combining all the steps above, we proved the theorem:

Theorem 10 (Properties of critical points)

The only critical points of $f(x)$ are of the form $x = \pm\sqrt{\lambda_i} v_i$ or $x = 0$. For all critical points except $x = \pm\sqrt{\lambda_1} v_1$, $\nabla^2 f(x)$ has a negative eigenvalue.

- The only 2^{nd}-order stationary points are $x = \pm\sqrt{\lambda_1} v_1$, so all 2^{nd}-order stationary points are also global minima.

Finding Directions of Improvements

It is often infeasible to enumerate all the solutions for $\nabla f(x) = 0$.
Key: For every point x that is not a global minimum, we define its direction of improvements as below:

Definition 11

— For an objective function f and a point x, we say δ is a direction of improvement (of f at x) if $|\langle \nabla f(x), \delta \rangle| > 0$ or $\delta^\top [\nabla^2 f(x)] \delta < 0$.
— We say δ is an (ϵ, γ)-direction of improvement $($of f at $x)$ if $|\langle \nabla f(x), \delta \rangle| > \epsilon \|\delta\|_2$ or $\delta^\top [\nabla^2 f(x)] \delta < -\gamma \|\delta\|_2^2$.

- If δ is a direction of improvement for f at x, then moving along one of δ or $-\delta$ for a small enough step can decrease the objective function.

Finding Directions of Improvements

For simplicity, consider an simpler version of the top eigenvector problem, where $M = zz^\top$ is a rank-1 matrix, and z is a unit vector. Then

$$\min_x f(x) = \frac{1}{4} \left\| zz^\top - xx^\top \right\|_F^2 \tag{4}$$

- Which direction should we move to decrease the objective?
- One only have the optimal direction z and the current direction x, so the natural guesses would be z, x or $z - x$.

Finding Directions of Improvements

Lemma 12

For objective function f, there exists a universal constant $c > 0$ such that for any $\tau < 1$, if neither x or z is an $(c\tau, 1/4)$-direction of improvement for the point x, then $f(x) \leqslant \tau$.

- The proof of this lemma involves some detailed calculation.
- To get some intuition, we first think about what happens if neither x or z is a direction of improvement.

Finding Directions of Improvements

Lemma 13

For objective function f, if neither x or z is a direction of improvement of f at x, then $f(x) = 0$.

Proof.

If x is not a direction of improvement, we must have:

$$\langle \nabla f(x), x \rangle = 0 \Longrightarrow \|x\|_2^4 = \langle x, z \rangle^2$$

If z is not a direction of improvement, we know $z^\top \left[\nabla^2 f(x)\right] z \geqslant 0$ which means

$$\|x\|^2 + 2\langle x, z \rangle^2 - 1 \geqslant 0 \Longrightarrow \|x\|^2 \geqslant 1/3$$

□

Finding Directions of Improvements

Proof.

Consider the fact that $\langle x, z\rangle^2 \leqslant \|x\|_2^2 \|z\|_2^2 = \|x\|_2^2$, thus we have $\langle x, z\rangle^2 = \|x\|_2^4 \geqslant 1/9$.

Finally, since z is not a direction of improvement, we know $\langle \nabla f(x), z\rangle = 0$, which implies $\langle x, z\rangle \left(\|x\|_2^2 - 1\right) = 0$. We have already proved $\langle x, z\rangle^2 \geqslant 1/9 > 0$, thus $\|x\|_2^2 = 1$.

Again we know $\langle x, z\rangle^2 = \|x\|_2^4 = 1$. The only two vectors with $\langle x, z\rangle^2 = 1$ and $\|x\|_2^2 = 1$ are $x = \pm z$. □

Finding Directions of Improvements

- The proof of Lemma 12 is very similar to Lemma 13, except we need to allow slacks in every equation and inequality we use.
- Lemma 12 and Lemma 13 both use directions x and z. It is also possible to use the direction $x - z$ when $\langle x, z \rangle \geqslant 0$ (and $x + z$ when $\langle x, z \rangle < 0$).
- Both ideas can be generalized to handle the case when $M = ZZ^\top$ where $Z \in \mathbb{R}^{d \times r}$, so M is a rank-r matrix.

Overview

Summary

- **Cases with a unique global minimum:** identify properties of the objective function, such as PL, weakly-quasi-convex and RSI conditions.
- **The permutation symmetry has many implication:** nonconvex, saddle points and so on.
- **Two strategies for analyzing the landscape:**
  - Characterizing all critical points
  - Finding directions of improvements

Summary

- **Cases with a unique global minimum:** identify properties of the objective function, such as PL, weakly-quasi-convex and RSI conditions.
- **The permutation symmetry has many implication:** nonconvex, saddle points and so on.
- **Two strategies for analyzing the landscape:**
  - Characterizing all critical points
  - Finding directions of improvements
- More latest studies showed:
  - Gradient descent converges at a global linear rate to the global optimum for wide two-layer NNs [Du et al., 2019].
  - How piecewise linear activation functions substantially shape the loss surfaces of neural networks [He et al., 2020].

References I

Du, S. S., Zhai, X., Póczos, B., and Singh, A. (2019).
Gradient descent provably optimizes over-parameterized neural networks.
In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ge, R., Huang, F., Jin, C., and Yuan, Y. (2015).
Escaping from saddle points—online stochastic gradient for tensor decomposition.
In *Conference on learning theory*, pages 797–842. PMLR.

Ge, R., Lee, J. D., and Ma, T. (2016).
Matrix completion has no spurious local minimum.
In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Ge, R., Lee, J. D., and Ma, T. (2017).
Learning one-hidden-layer neural networks with landscape design.
arXiv preprint arXiv:1711.00501.

Hardt, M., Ma, T., and Recht, B. (2016).
Gradient descent learns linear dynamical systems.
arXiv preprint arXiv:1609.05191.

References II

He, F., Wang, B., and Tao, D. (2020).
Piecewise linear activations substantially shape the loss surfaces of neural networks.
In *International Conference on Learning Representations*.

Kakade, S. M., Kanade, V., Shamir, O., and Kalai, A. (2011).
Efficient learning of generalized linear and single index models with isotonic regression.
In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Kalai, A. T. and Sastry, R. (2009).
The isotron algorithm: High-dimensional isotonic regression.
In *COLT*. Citeseer.

Sun, J., Qu, Q., and Wright, J. (2016).
Complete dictionary recovery over the sphere i: Overview and the geometric picture.
IEEE Transactions on Information Theory, 63(2):853–884.