Estimation of Extreme Values and Associated Level Sets of a Regression Function via Selective Sampling

Stanislav Minsker

Duke University

Estimation framework

- Let (X, Y) be a random couple in $[0,1]^d \times \mathbb{R}$:

$$Y = f(X) + \xi,$$
$$f(x) = \mathbb{E}(Y|X = x)$$

- The joint distribution of (X, Y) will be denoted by P and the distribution of X by Π.
- Goal: estimate

$$M(f) := \sup\{f(x),\ x \in \mathrm{supp}(\Pi)\}$$

and

$$L_M := \{x \in \mathrm{supp}(\Pi) :\ f(x) = M(f)\}$$

given a collection of noisy values $(X_i, Y_i),\ i = 1 \ldots n$.

Estimation framework

- Let (X, Y) be a random couple in $[0,1]^d \times \mathbb{R}$:

$$Y = f(X) + \xi,$$
$$f(x) = \mathbb{E}(Y|X = x)$$

- The joint distribution of (X, Y) will be denoted by P and the distribution of X by Π.
- Goal: estimate

$$M(f) := \sup\{f(x),\ x \in \mathrm{supp}(\Pi)\}$$

and

$$L_M := \{x \in \mathrm{supp}(\Pi) :\ f(x) = M(f)\}$$

given a collection of noisy values $(X_i, Y_i),\ i = 1 \ldots n$.

Estimation framework

- Let (X, Y) be a random couple in $[0,1]^d \times \mathbb{R}$:

$$Y = f(X) + \xi,$$
$$f(x) = \mathbb{E}(Y|X = x)$$

- The joint distribution of (X, Y) will be denoted by P and the distribution of X by Π.
- Goal: estimate

$$M(f) := \sup\{f(x),\ x \in \mathrm{supp}(\Pi)\}$$

and

$$L_M := \{x \in \mathrm{supp}(\Pi) :\ f(x) = M(f)\}$$

given a collection of noisy values $(X_i, Y_i),\ i = 1 \ldots n$.

“Passive" and “Active" frameworks

- In the **passive learning** framework the collection $(X_i, Y_i)_{i=1}^n$ is a given input.
- Not very efficient: many of the data points are not informative.
- Active algorithms try to select the most informative locations.

“Passive" and “Active" frameworks

- In the passive learning framework the collection $(X_i, Y_i)_{i=1}^n$ is a given input.
- Not very efficient: many of the data points are not informative.
- Active algorithms try to select the most informative locations.

“Passive" and “Active" frameworks

- In the passive learning framework the collection $(X_i, Y_i)_{i=1}^n$ is a given input.
- Not very efficient: many of the data points are not informative.
- Active algorithms try to select the most informative locations.

“Passive" and “Active" frameworks

Sampling models:

1. Observations are sampled sequentially.
   Support $\mathrm{supp}(\Pi)$ is known and the algorithm is allowed to evaluate f at *any* $X_k \in \mathrm{supp}(\Pi)$.
2. Streaming mode: accept-reject.
   X_k is sampled from the modified distribution $\hat{\Pi}_k = \Pi(\cdot|\hat{A}_k)$
   where $\hat{A}_k$ depends on $(X_1, Y_1), \ldots, (X_{k-1}, Y_{k-1})$.

   - Y_k is sampled from the conditional distribution $P_{Y|X}(\cdot|X = x)$.
     $\{Y_k\}$ are conditionally independent given the feature vectors $\{X_i\}$.

“Passive" and “Active" frameworks

Sampling models:

1. Observations are sampled sequentially.
   Support $\mathrm{supp}(\Pi)$ is known and the algorithm is allowed to evaluate f at *any* $X_k \in \mathrm{supp}(\Pi)$.
2. Streaming mode: accept-reject.
   X_k is sampled from the modified distribution $\hat{\Pi}_k = \Pi(\cdot|\hat{A}_k)$
   where $\hat{A}_k$ depends on $(X_1, Y_1), \ldots, (X_{k-1}, Y_{k-1})$.

- Y_k is sampled from the conditional distribution $P_{Y|X}(\cdot|X = x)$.
  $\{Y_k\}$ are conditionally independent given the feature vectors $\{X_i\}$.

“Passive" and “Active" frameworks

Sampling models:

1. Observations are sampled sequentially.
   Support $\mathrm{supp}(\Pi)$ is known and the algorithm is allowed to evaluate f at *any* $X_k \in \mathrm{supp}(\Pi)$.
2. Streaming mode: accept-reject.
   X_k is sampled from the modified distribution $\hat{\Pi}_k = \Pi(\cdot|\hat{A}_k)$
   where $\hat{A}_k$ depends on $(X_1, Y_1), \ldots, (X_{k-1}, Y_{k-1})$.

- Y_k is sampled from the conditional distribution $P_{Y|X}(\cdot|X = x)$.
  $\{Y_k\}$ are conditionally independent given the feature vectors $\{X_i\}$.

Previous work

1. J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function (1952).
2. R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandits in metric spaces (2008).
3. S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvàri. $\mathfrak{X}$-armed bandits (2011).
4. E. Belitser, S. Ghosal, and H. van Zanten. Optimal two-stage procedures for estimating location and size of maximum of multivariate regression functions (2012).
5. V. Perchet, V. and P. Rigollet. The multi-armed bandit problem with covariates (2013).

. . . . . .

Previous work

Important parameters of the problem:

1. Regularity of the function:

$$|f(x) - f(y)| \leq K\|x - y\|^{\beta}, \ \beta \in (0, 1].$$

2. Margin condition (closely related to “Zooming dimension"): for all $t > 0$,

$$\Pi\{x : \ M(f) - f(x) \leq t\} \leq Bt^{\gamma}, \ \gamma > 0.$$

Previous work

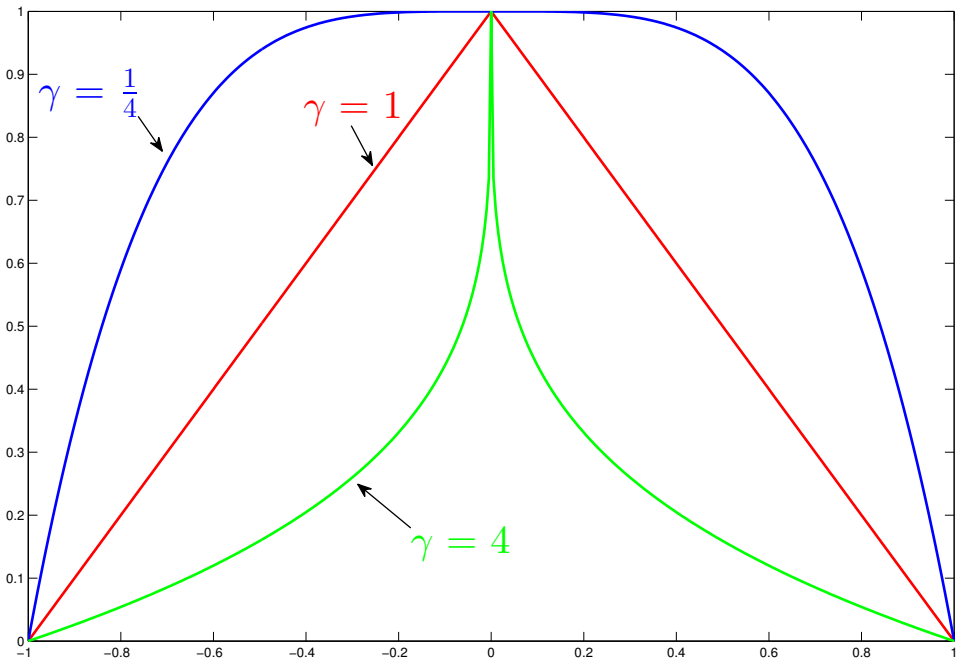
Important parameters of the problem:

1. Regularity of the function:

$$|f(x) - f(y)| \leq K\|x - y\|^{\beta}, \ \beta \in (0, 1].$$

2. Margin condition (closely related to "Zooming dimension"): for all $t > 0$,

$$\Pi\{x : \ M(f) - f(x) \leq t\} \leq Bt^{\gamma}, \ \gamma > 0.$$



Previous work

Important parameters of the problem:

1. Regularity of the function:

$$|f(x) - f(y)| \leq K\|x - y\|^{\beta}, \ \beta \in (0, 1].$$

2. Margin condition (closely related to “Zooming dimension"): for all $t > 0$,

$$\Pi\{x : \ M(f) - f(x) \leq t\} \leq Bt^{\gamma}, \ \gamma > 0.$$

Theorem [P. Auer, R. Ortner, and C. Szepesvàri (2007)]

In the univariate case, for any estimator $\hat{M}_n = \hat{M}_n(X_i, Y_i)$ of $M(f)$ and any $\varepsilon > 0$

$$\limsup_{n\to\infty} \frac{\mathbb{E}|\hat{M}_n - M(f)|}{n^{-\frac{\beta+\varepsilon}{2\beta+1-\beta\gamma}}} = \infty$$

Motivation

- Develop algorithms that require minimal assumptions and are adaptive (with respect to smoothness and margin parameters).
- Support of Π might be unknown (in other words, we are looking at the problem with unknown constraints).
  In particular, $\mathrm{supp}(\Pi)$ can be an m-dimensional submanifold of $\mathbb{R}^d$, where $m \ll d$.
  Can we take advantage of such structure?

Motivation

- Develop algorithms that require minimal assumptions and are adaptive (with respect to smoothness and margin parameters).
- Support of Π might be unknown (in other words, we are looking at the problem with unknown constraints).
  In particular, $\mathrm{supp}(\Pi)$ can be an m-dimensional submanifold of $\mathbb{R}^d$, where $m \ll d$.
  Can we take advantage of such structure?

Motivation

- Develop algorithms that require minimal assumptions and are adaptive (with respect to smoothness and margin parameters).
- Support of Π might be unknown (in other words, we are looking at the problem with unknown constraints).
  In particular, $\mathrm{supp}(\Pi)$ can be an m-dimensional submanifold of $\mathbb{R}^d$, where $m \ll d$.
  Can we take advantage of such structure?
  Addressed in numerical simulation section.

Main results: Lower bound

Recall that the Hausdorff distance between non-empty sets $A, B \in \mathbb{R}^d$ is

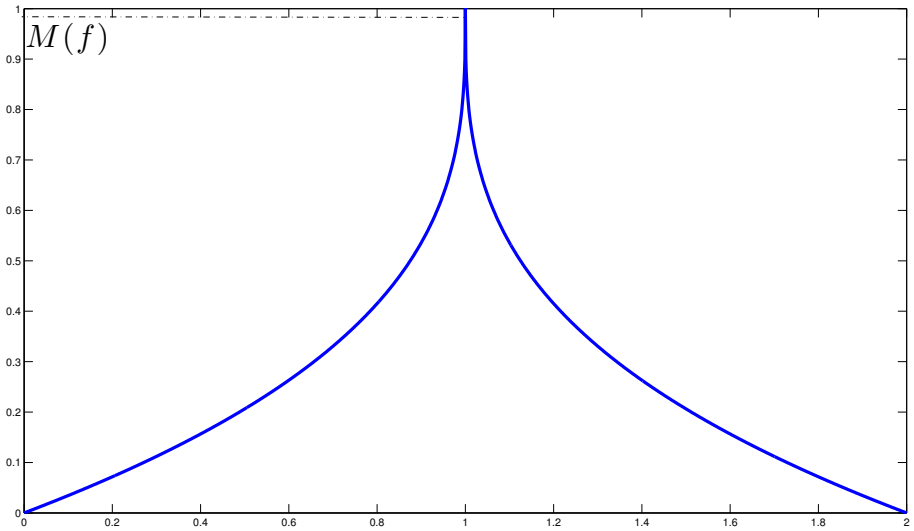
$$d_H(A,B) = \max\left(\sup_{x\in A}\inf_{y\in B}\|x-y\|_2, \sup_{y\in B}\inf_{x\in A}\|x-y\|_2\right),$$

Theorem (M., 2013)

Let $\beta \in (0,1]$, $\gamma > 0$. There exists $C > 0$ such that for all n large enough and for any estimator $\hat{L}_M$ based on n noisy measurements, we have

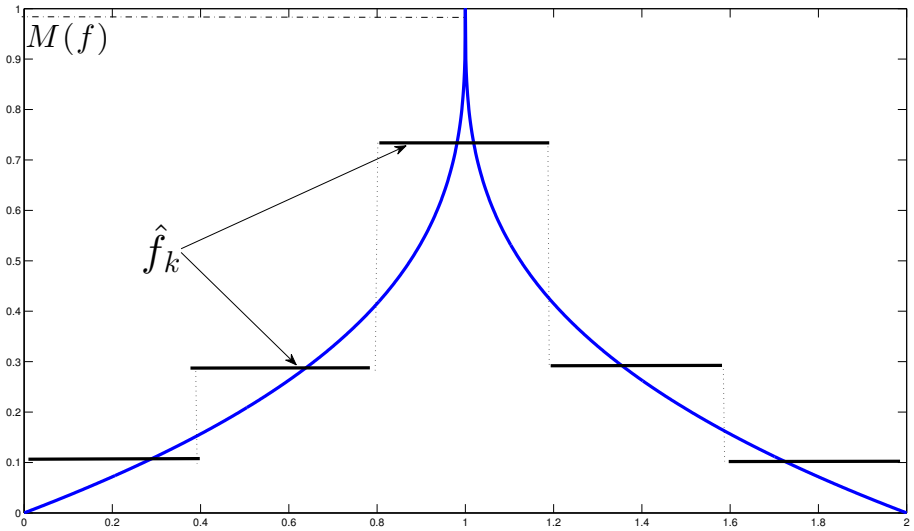
$$\mathbb{E}_P\, d_H\left(\hat{L}_M, L_M\right) \geq cn^{-\frac{1}{2\beta+d-\beta\gamma}}.$$

Description of an Algorithm



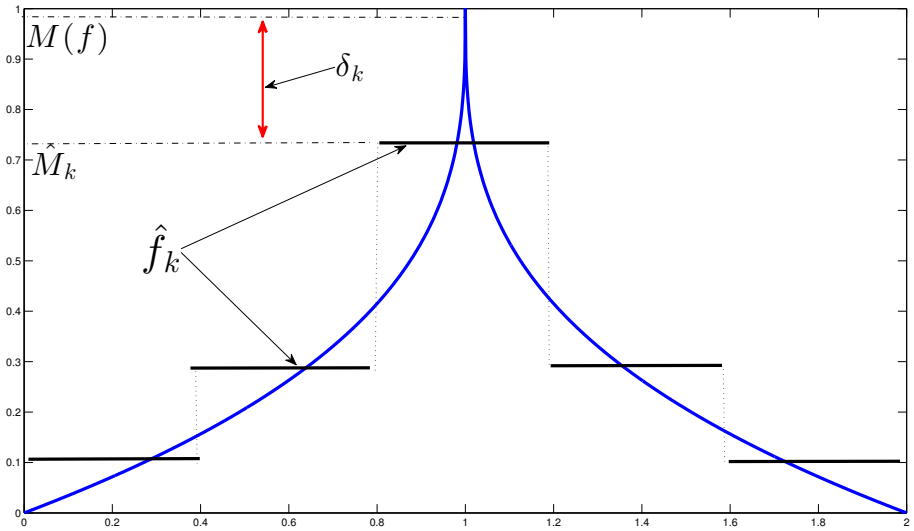
Suppose that regularity and margin assumptions are satisfied.

Description of an Algorithm



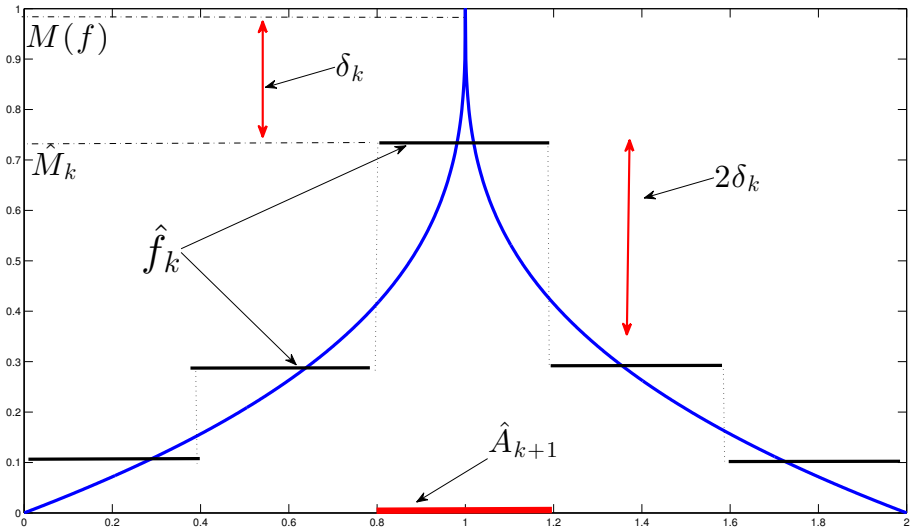
Use a small fraction of data to construct an estimator that is close to $f(x)$ in sup-norm. In our case, this is a piecewise–constant estimator.

Description of an Algorithm



- Assume that $\|\hat{f}_k - f\|_\infty \leq \delta_k$ with high probability.
- Define $\hat{M}_k = \sup \hat{f}_k$.

Description of an Algorithm

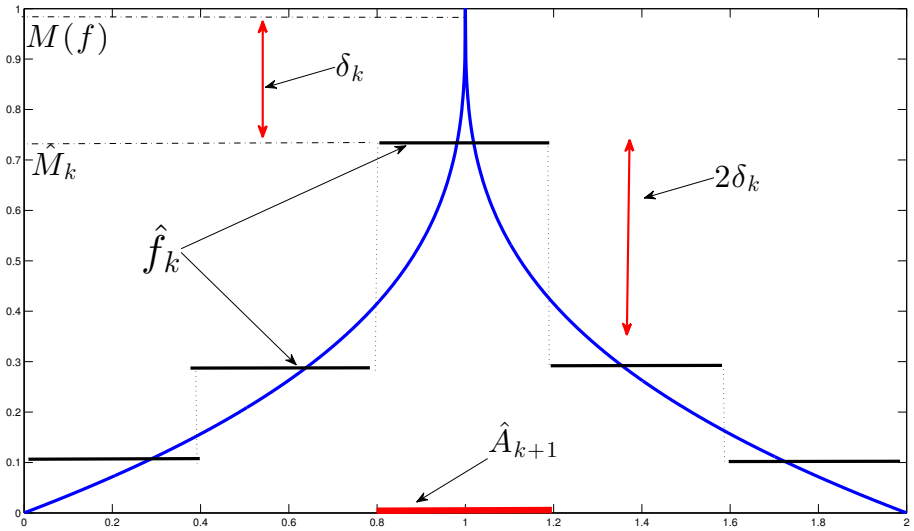


Define

$$\widehat{A}_{k+1} := \left\{ x : \ \hat{M}_k - \hat{\eta}_k(x) \leq 2\delta_k \right\}$$

With high probability, there are no candidates for the maxima outside of $\widehat{A}_{k+1}$.

Description of an Algorithm



Next Iteration:

- Sample new locations from $\widehat{A}_{k+1}$.
- Construct a tighter confidence band;
- Repeat until the data limit is reached;
- Return $\hat{M}_L$ and $\widehat{A}_{L+1}$.

Theoretical guarantees: rates of convergence

Assumptions:

- Hölder regularity (for some unknown $\beta \in (\nu, 1]$);
- Margin conditions (for some unknown $\gamma > 0$);
- Noise ξ is bounded, or has subexponential tails and is independent of X.
- $\mathrm{supp}(\Pi) = [0,1]^d$, $d\Pi(x) = p(x)dx$, $0 < c_1 \leq p(x) \leq c_2 < \infty$ for all $x \in \mathrm{supp}(\Pi)$.

Theoretical guarantees: rates of convergence

Assumptions:

Key assumption for adaptivity: let f_m be the $L_2(\Pi)$ projection of f onto the linear span of first 2^{dm} Haar basis functions. Then

$$\|f - f_m\|_\infty \geq B_2 2^{-\beta m}$$

Theoretical guarantees: rates of convergence

Assumptions:

Key assumption for adaptivity: let f_m be the $L_2(\Pi)$ projection of f onto the linear span of first 2^{dm} Haar basis functions. Then

$$\underbrace{B_1 2^{-\beta m} \geq}_{\text{Hölder regularity}} \|f - f_m\|_\infty \geq B_2 2^{-\beta m}$$

Intuition: smoothness can be learned from the data.

Theoretical guarantees: rates of convergence

Assumptions:

Key assumption for adaptivity: let f_m be the $L_2(\Pi)$ projection of f onto the linear span of first 2^{dm} Haar basis functions. Then

$$\underbrace{B_1 2^{-\beta m} \geq}_{\text{Hölder regularity}} \|f - f_m\|_\infty \geq B_2 2^{-\beta m}$$

Intuition: smoothness can be learned from the data.

Theorem (M., 2013)

With probability $\geq 1 - \alpha$, estimators $\hat{L}_M$ and $\hat{M}_n$ returned by the Algorithm satisfy

$$L_M \subseteq \hat{L}_M \subseteq \left\{ x \in [0,1]^d : \ \eta(x) \geq M(\eta) - 4\varepsilon \right\},$$

$$|\hat{M} - M(\eta)| \leq \varepsilon,$$

while the total number of noisy function measurements requested by the Algorithm is

$$n \leq C \left(\frac{1}{\varepsilon}\right)^{\frac{2\beta + d - \beta\gamma}{\beta}} \operatorname{polylog}\left(\frac{1}{\varepsilon\alpha}\right).$$

Comments

- "Key assumption" is satisfied for most "nice" (real-world) functions.
  For example, all continuously differentiable functions satisfy it for $\beta = 1$.
- "Zero-order" algorithm: unable to take full advantage of higher order smoothness.
  For smooth functions, the sample complexity can be bounded as

$$n \leq C \left(\frac{1}{\varepsilon}\right)^{\frac{2\beta+d-(\beta\wedge 1)\gamma}{\beta}} \operatorname{polylog}\left(\frac{1}{\varepsilon\alpha}\right).$$

- $\operatorname{supp}(\Pi)$ is assumed to be known. Would like to avoid this in practice.
- It is possible to use similar ideas to develop an implementable and computationally efficient online algorithm.

Comments

- “Key assumption" is satisfied for most “nice" (real-world) functions.
  For example, all continuously differentiable functions satisfy it for $\beta = 1$.
- “Zero-order" algorithm: unable to take full advantage of higher order smoothness.
  For smooth functions, the sample complexity can be bounded as

$$n \leq C \left(\frac{1}{\varepsilon}\right)^{\frac{2\beta+d-(\beta\wedge 1)\gamma}{\beta}} \operatorname{polylog}\left(\frac{1}{\varepsilon\alpha}\right).$$

- $\operatorname{supp}(\Pi)$ is assumed to be known. Would like to avoid this in practice.
- It is possible to use similar ideas to develop an implementable and computationally efficient online algorithm.

Comments

- “Key assumption" is satisfied for most “nice" (real-world) functions.
  For example, all continuously differentiable functions satisfy it for $\beta = 1$.
- “Zero-order" algorithm: unable to take full advantage of higher order smoothness.
  For smooth functions, the sample complexity can be bounded as

$$n \leq C \left(\frac{1}{\varepsilon}\right)^{\frac{2\beta+d-(\beta\wedge 1)\gamma}{\beta}} \operatorname{polylog}\left(\frac{1}{\varepsilon\alpha}\right).$$

- $\operatorname{supp}(\Pi)$ is assumed to be known. Would like to avoid this in practice.
- It is possible to use similar ideas to develop an implementable and computationally efficient online algorithm.

Comments

- “Key assumption" is satisfied for most “nice" (real-world) functions.
  For example, all continuously differentiable functions satisfy it for $\beta = 1$.
- “Zero-order" algorithm: unable to take full advantage of higher order smoothness.
  For smooth functions, the sample complexity can be bounded as

$$n \leq C \left(\frac{1}{\varepsilon}\right)^{\frac{2\beta+d-(\beta\wedge 1)\gamma}{\beta}} \operatorname{polylog}\left(\frac{1}{\varepsilon\alpha}\right).$$

- $\operatorname{supp}(\Pi)$ is assumed to be known. Would like to avoid this in practice.
- It is possible to use similar ideas to develop an implementable and computationally efficient online algorithm.

Simulation

Model:

$$Y = f(X) + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2) \text{ and is independent of } X$$

$$f(X) = -\|x_0 - X\|_2^{\gamma},\ x_0 \in \operatorname{supp}(\Pi)$$

1. $\Pi \sim$ Uniform $\left([-1,1]^3\right)$, $x_0 = (0,0,0)$, $\gamma = \frac{3}{8}$.
2. $\Pi \sim$ Uniform $\left(S^2\right)$, $x_0 = (0,0,1)$, $\gamma = \frac{3}{8}$.

$\operatorname{supp}(\Pi)$ is not given to an algorithm

Simulation

Model:

$$Y = f(X) + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2) \text{ and is independent of } X$$

$$f(X) = -\|x_0 - X\|_2^{\gamma},\ x_0 \in \operatorname{supp}(\Pi)$$

1. $\Pi \sim$ Uniform $\left([-1,1]^3\right)$, $x_0 = (0,0,0)$, $\gamma = \frac{3}{8}$.
2. $\Pi \sim$ Uniform $\left(S^2\right)$, $x_0 = (0,0,1)$, $\gamma = \frac{3}{8}$.

$\operatorname{supp}(\Pi)$ is not given to an algorithm

Method:

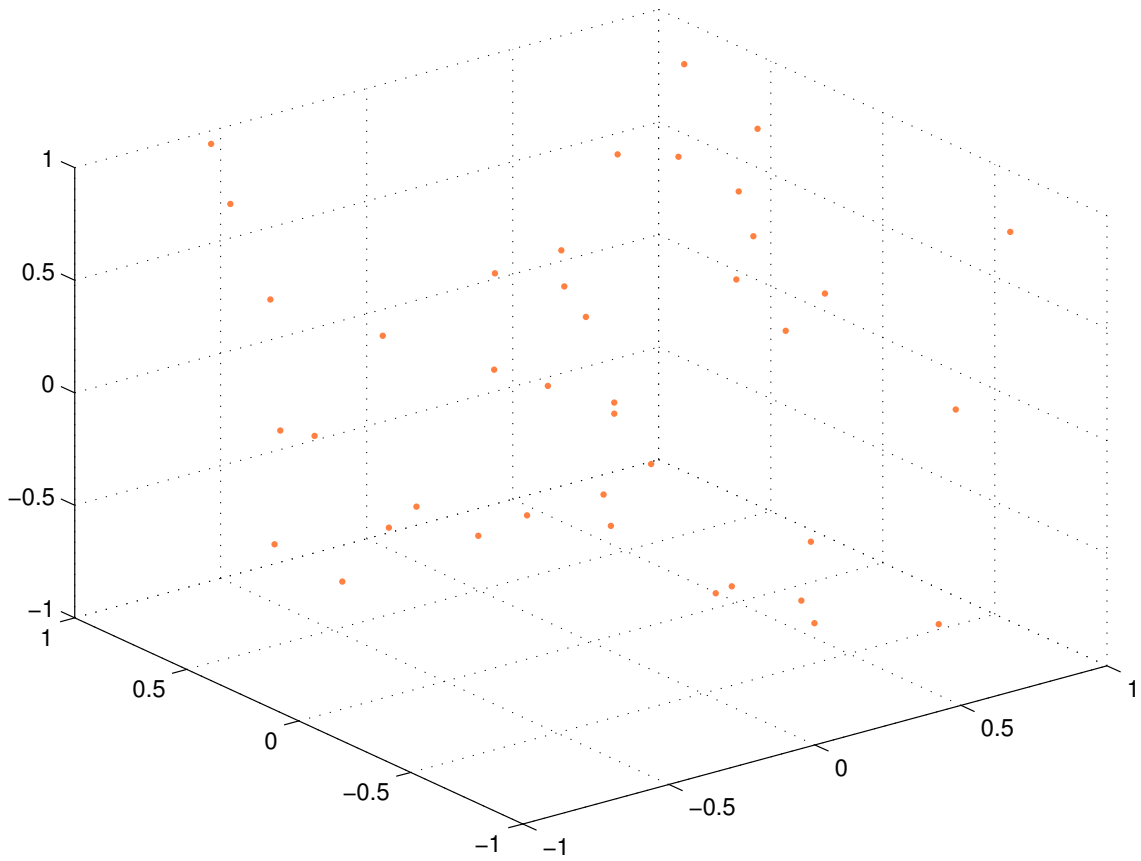
- Similar to described algorithm, but evaluates $\hat{f}_k(\cdot)$ only at observed locations.
- Uses more flexible estimators.

Scenario 1

$$f(X) = -\|X\|_2^{3/8}$$
$$\Pi \sim \text{Uniform}\left([-1,1]^3\right)$$
$$\sigma = 0.1$$

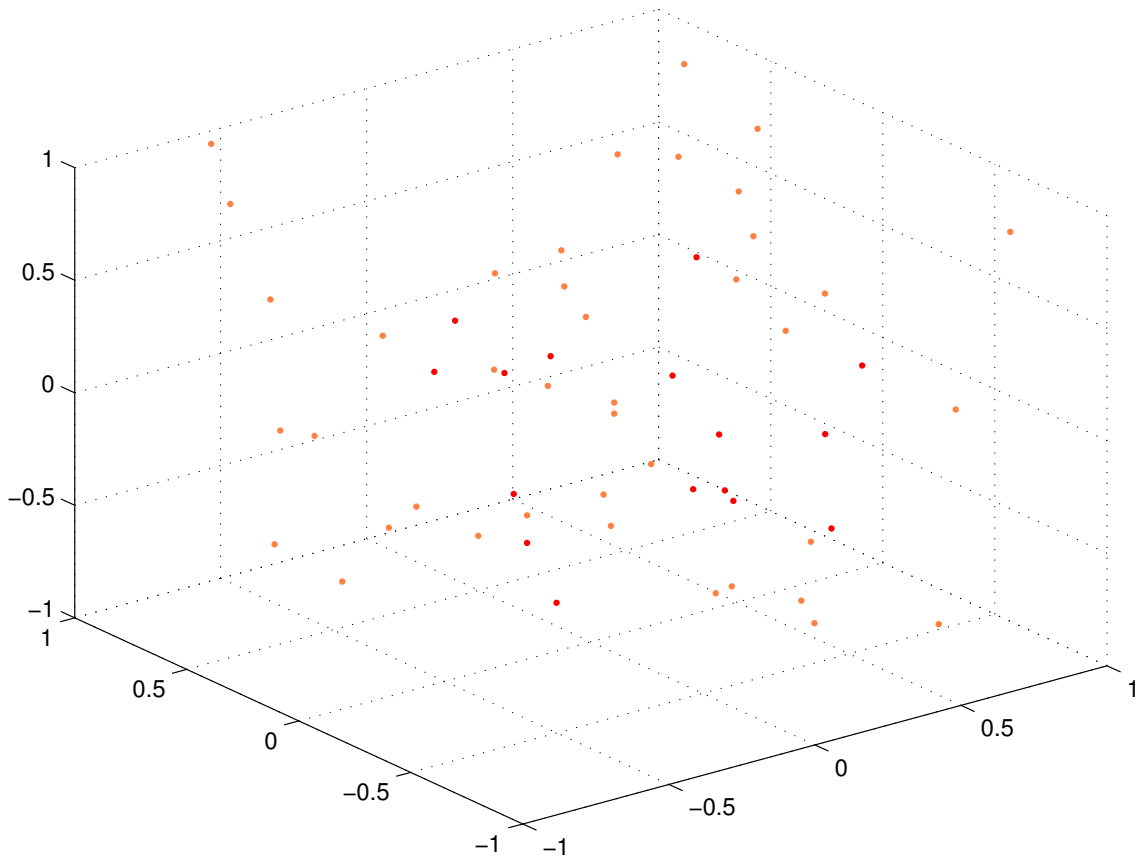
Scenario 1

$$f(X) = -\|X\|_2^{3/8}$$
$$\Pi \sim \text{Uniform}\left([-1,1]^3\right)$$
$$\sigma = 0.1$$



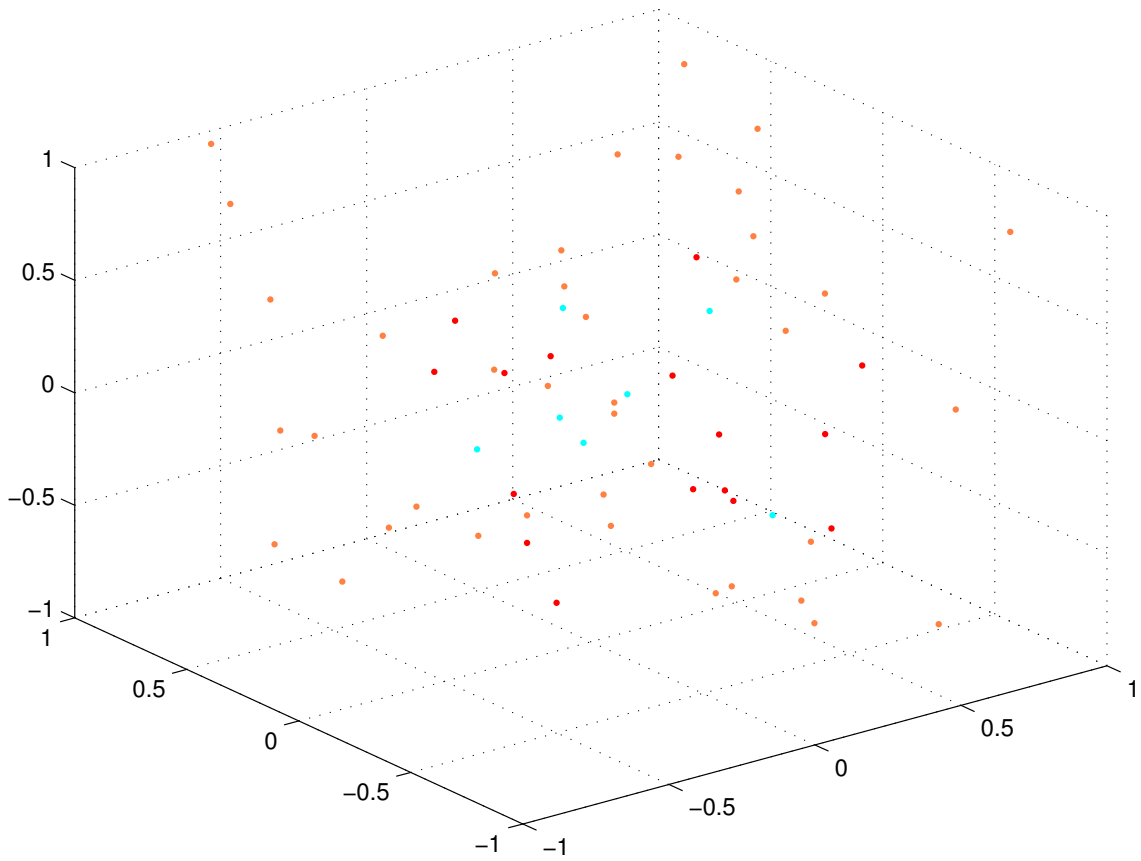
Scenario 1

$$f(X) = -\|X\|_2^{3/8}$$
$$\Pi \sim \text{Uniform}\left([-1,1]^3\right)$$
$$\sigma = 0.1$$



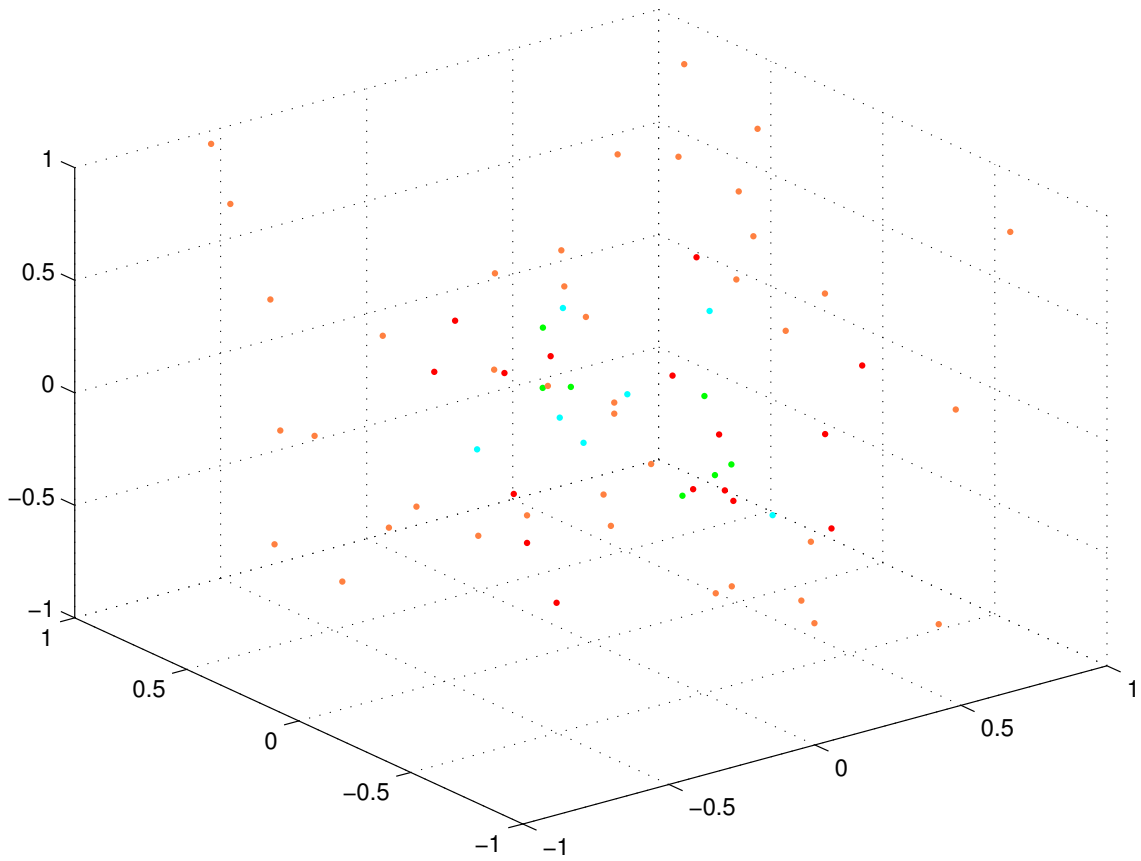
Scenario 1

$$f(X) = -\|X\|_2^{3/8}$$
$$\Pi \sim \text{Uniform}\left([-1,1]^3\right)$$
$$\sigma = 0.1$$



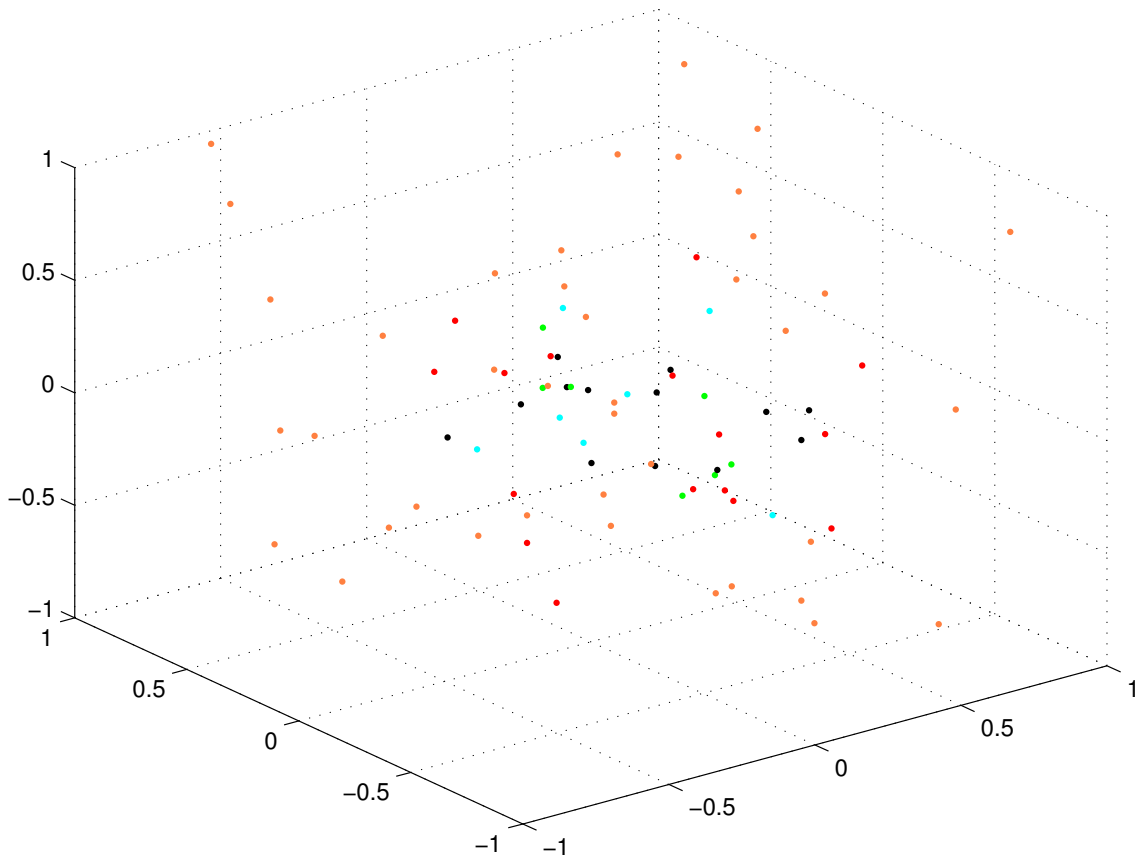
Scenario 1

$$f(X) = -\|X\|_2^{3/8}$$
$$\Pi \sim \text{Uniform}\left([-1,1]^3\right)$$
$$\sigma = 0.1$$



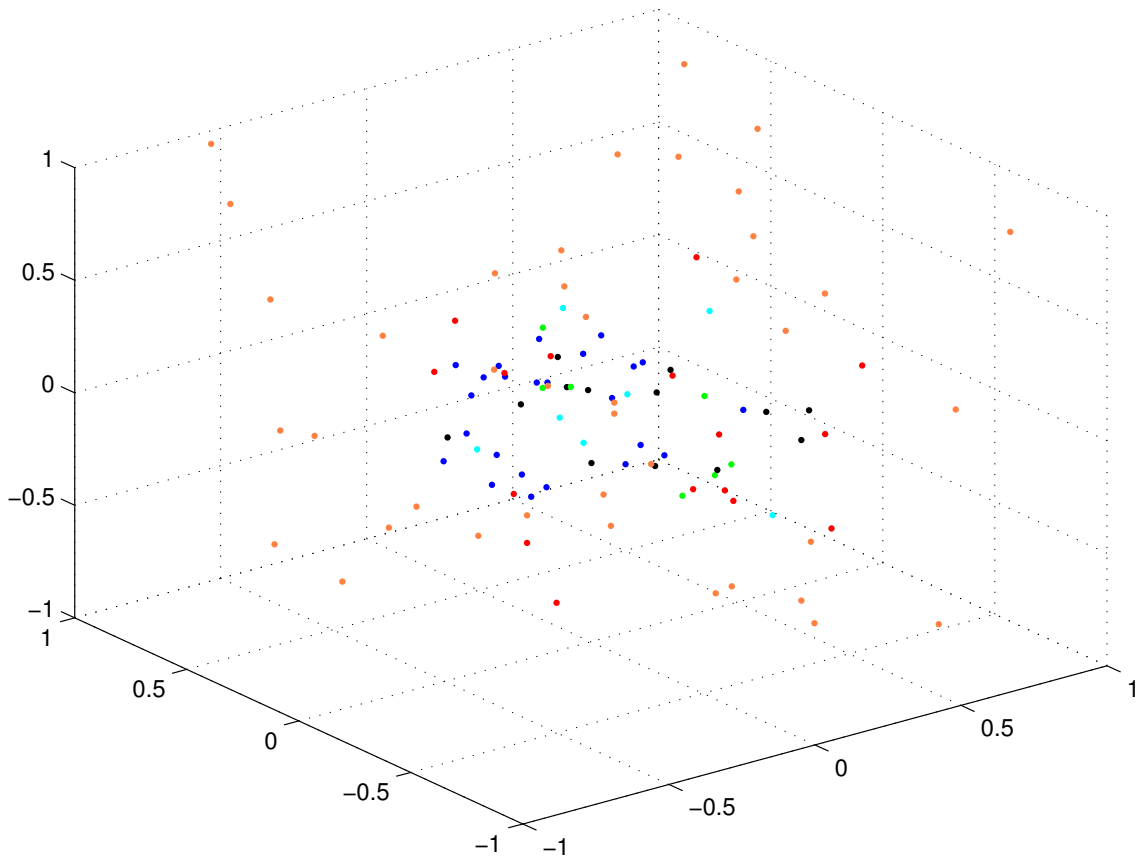
Scenario 1

$$f(X) = -\|X\|_2^{3/8}$$
$$\Pi \sim \text{Uniform}\left([-1,1]^3\right)$$
$$\sigma = 0.1$$



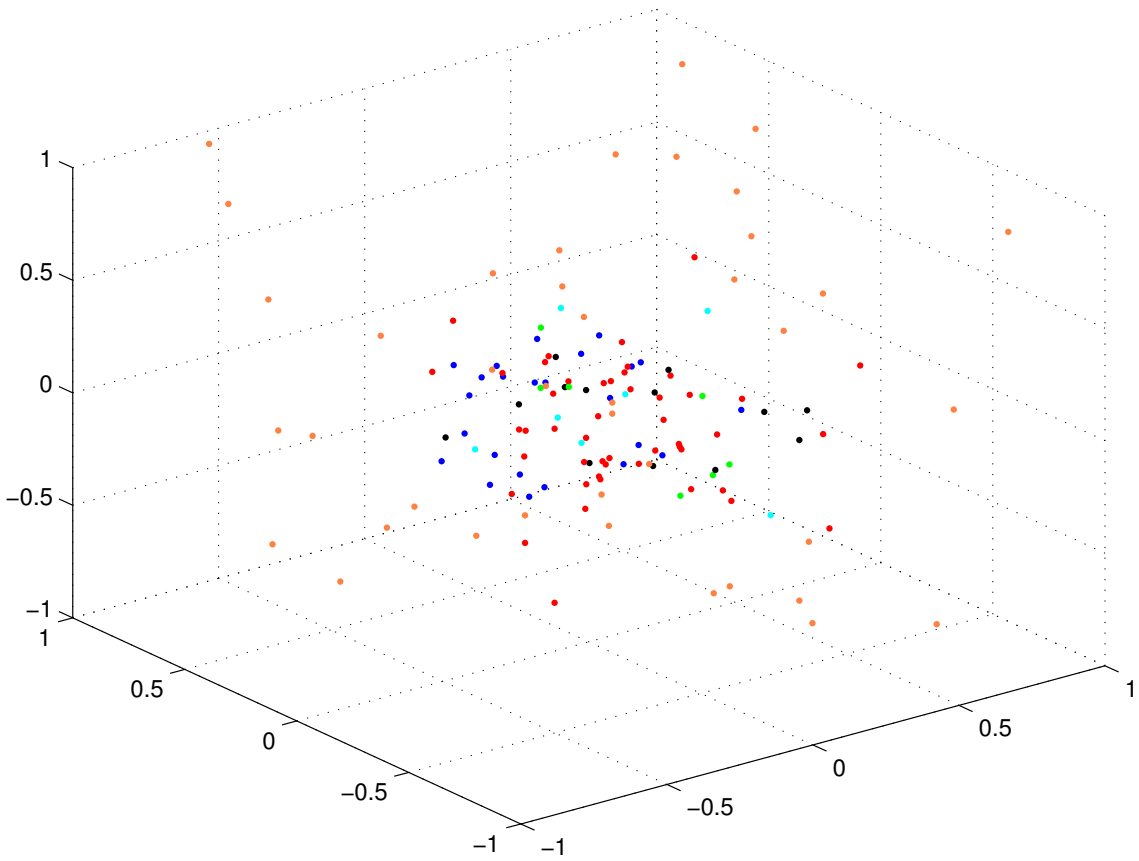
Scenario 1

$$f(X) = -\|X\|_2^{3/8}$$
$$\Pi \sim \text{Uniform}\left([-1,1]^3\right)$$
$$\sigma = 0.1$$



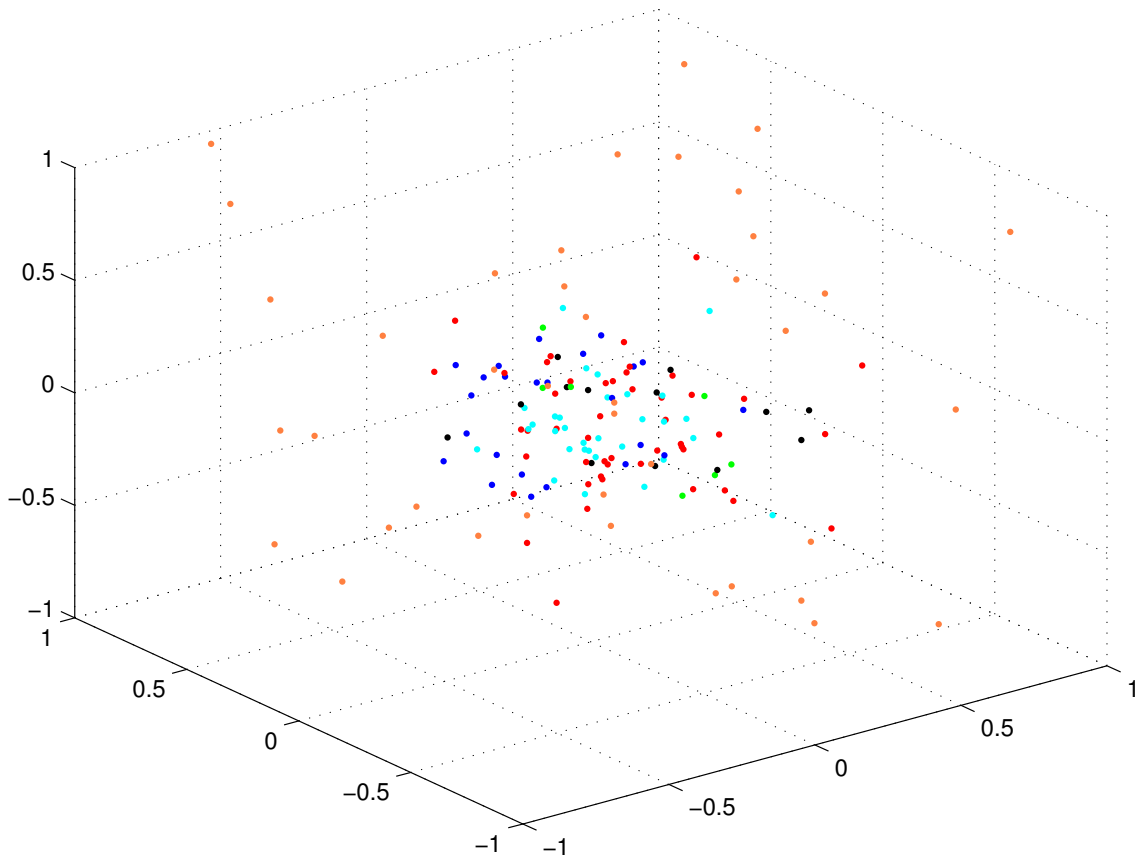
Scenario 1

$$f(X) = -\|X\|_2^{3/8}$$
$$\Pi \sim \text{Uniform}\left([-1,1]^3\right)$$
$$\sigma = 0.1$$



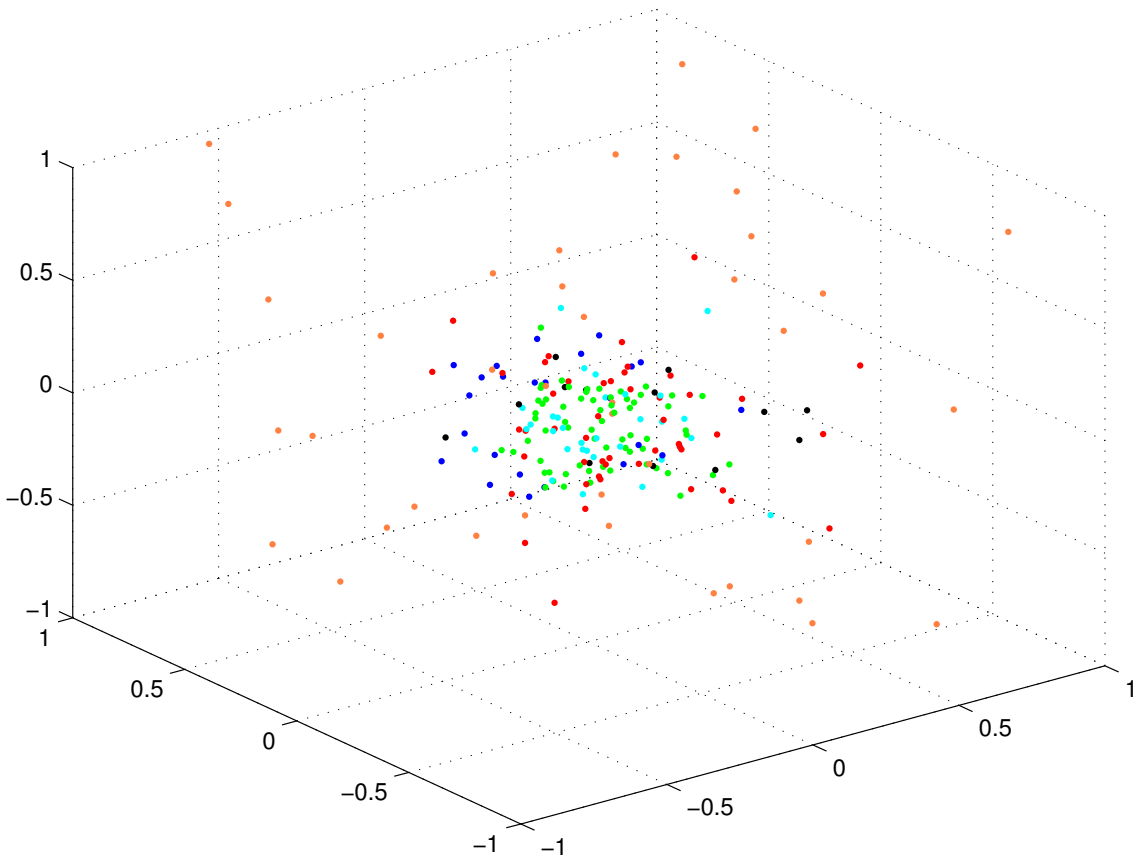
Scenario 1

$$f(X) = -\|X\|_2^{3/8}$$
$$\Pi \sim \text{Uniform}\left([-1,1]^3\right)$$
$$\sigma = 0.1$$



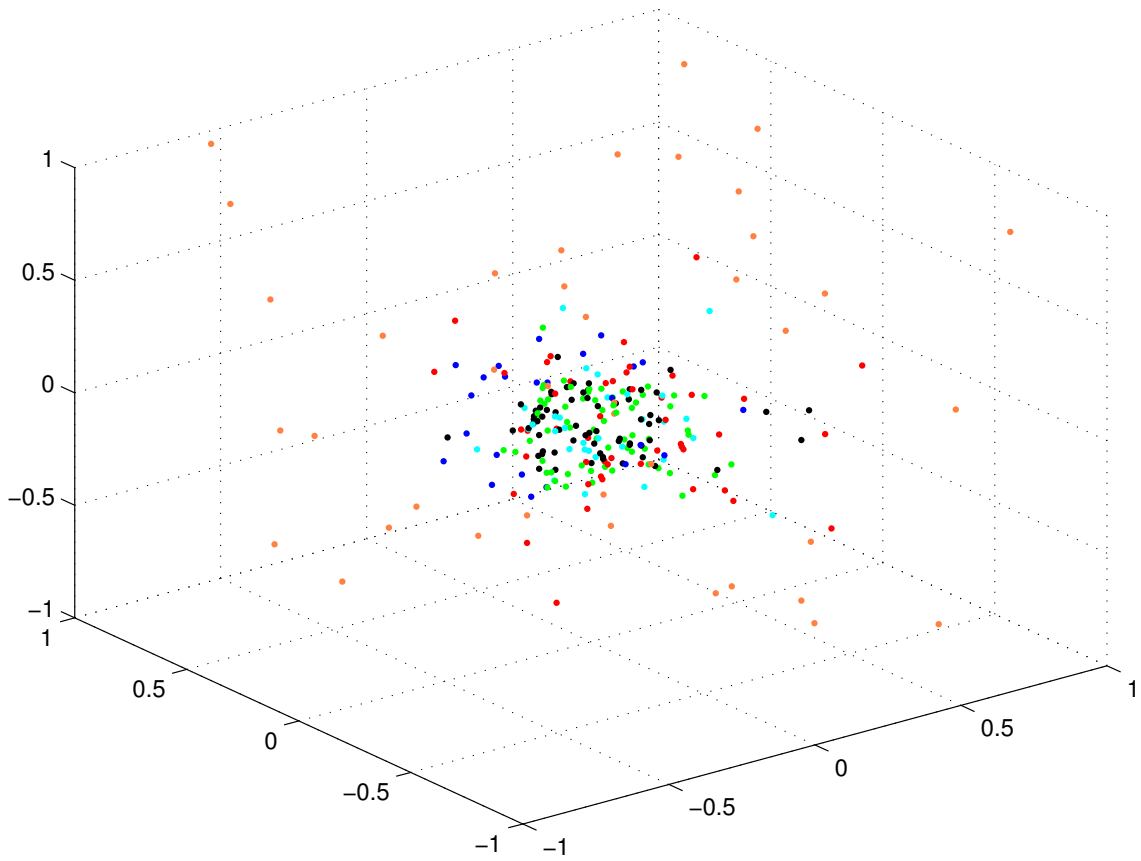
Scenario 1

$$f(X) = -\|X\|_2^{3/8}$$
$$\Pi \sim \text{Uniform}\left([-1,1]^3\right)$$
$$\sigma = 0.1$$



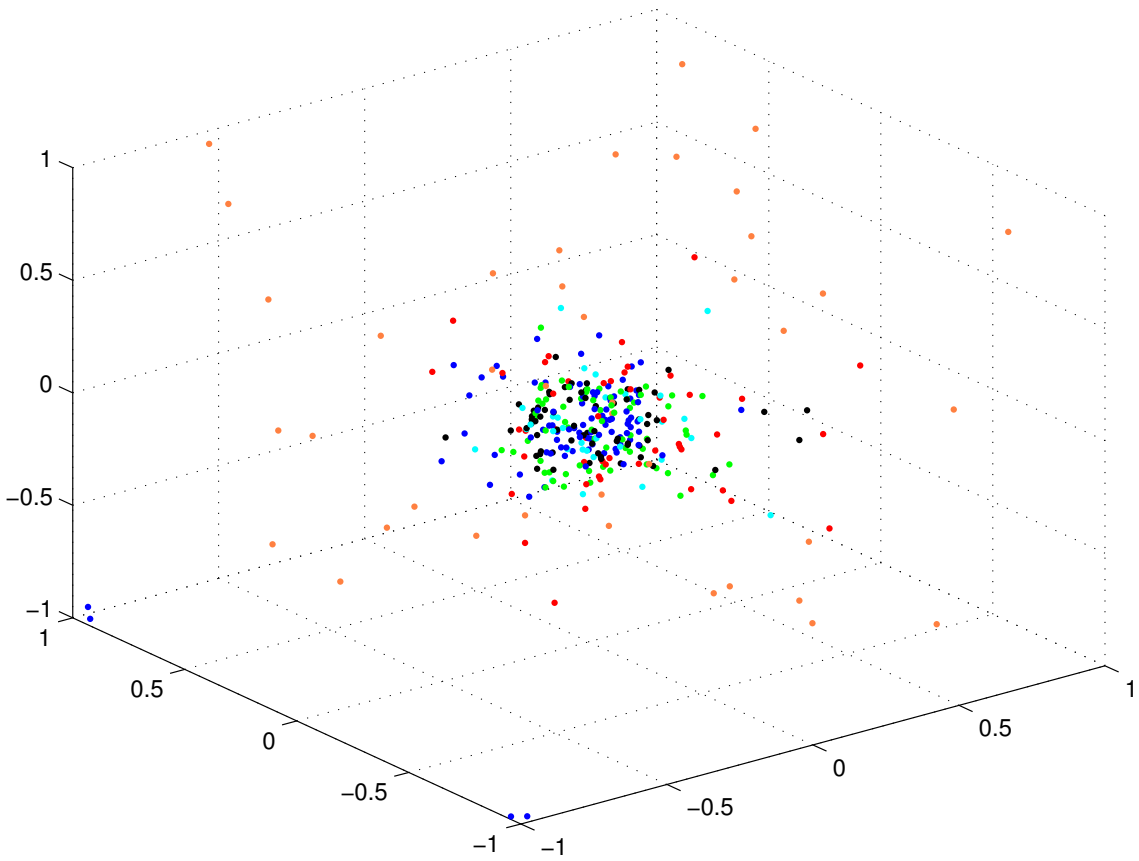
Scenario 1

$$f(X) = -\|X\|_2^{3/8}$$
$$\Pi \sim \text{Uniform}\left([-1,1]^3\right)$$
$$\sigma = 0.1$$



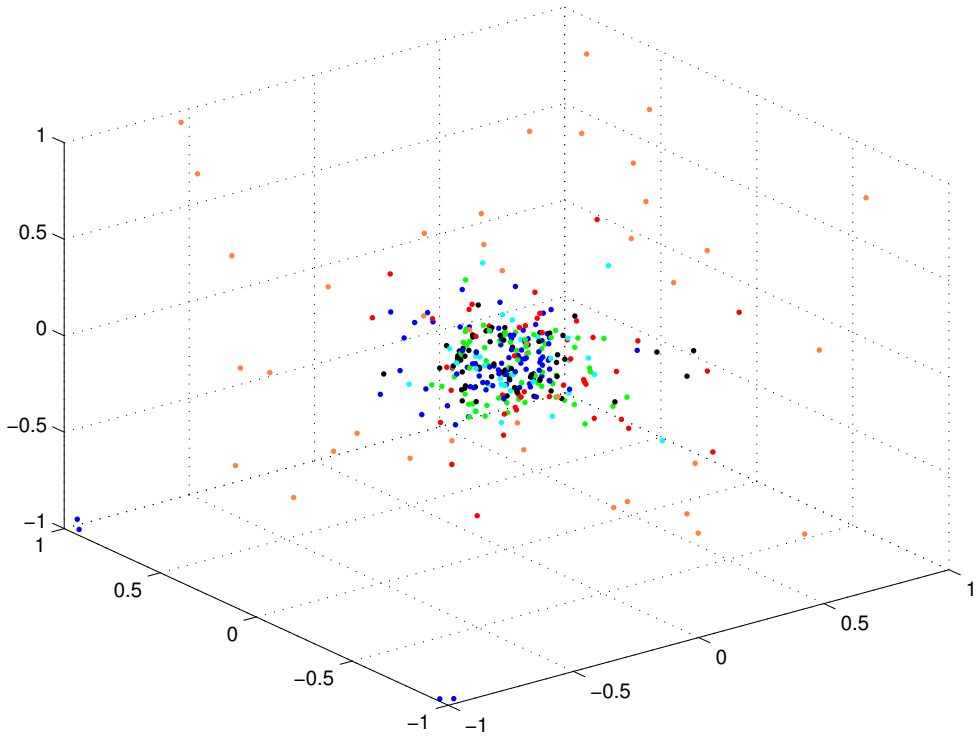
Scenario 1

$$f(X) = -\|X\|_2^{3/8}$$
$$\Pi \sim \text{Uniform}\left([-1,1]^3\right)$$
$$\sigma = 0.1$$



Scenario 1

$$f(X) = -\|X\|_2^{3/8}$$
$$\Pi \sim \text{Uniform}\left([-1,1]^3\right)$$
$$\sigma = 0.1$$



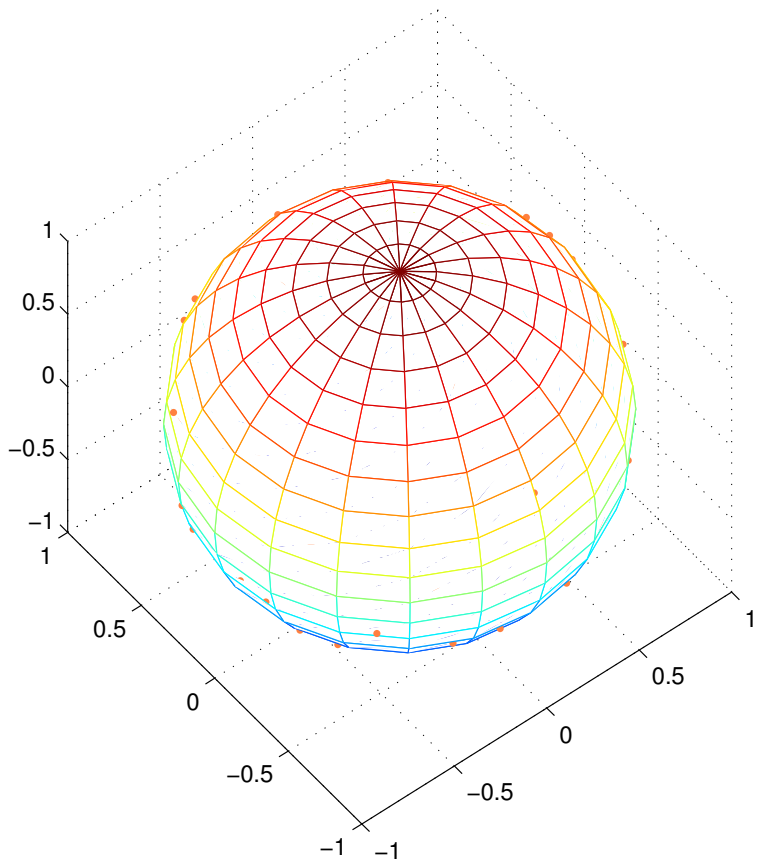
Estimated values:
argmax $= (-0.003, -0.021, 0.02)$
max $= -0.1896$
# of function evaluations $= 340$
Total # of processed samples $= 40960$.

Scenario 2

$$f(X) = -\|(0,0,1) - X\|_2^{3/4}$$
$$\Pi \sim \text{Uniform}\left(S^2\right)$$
$$\sigma = 0.3$$

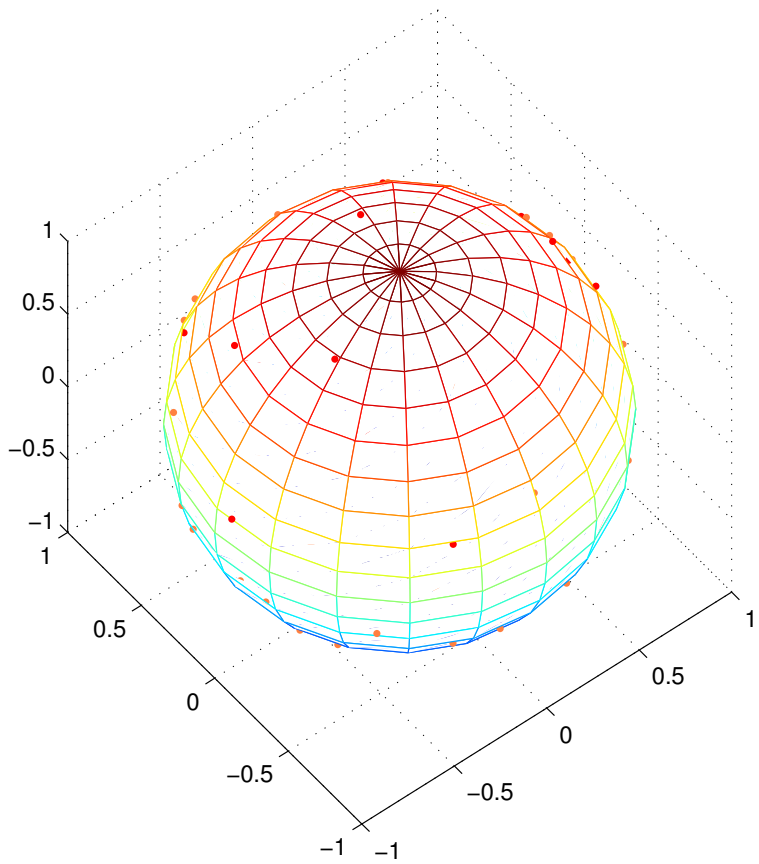
Scenario 2

$$f(X) = -\|(0,0,1) - X\|_2^{3/4}$$
$$\Pi \sim \text{Uniform}\left(S^2\right)$$
$$\sigma = 0.3$$



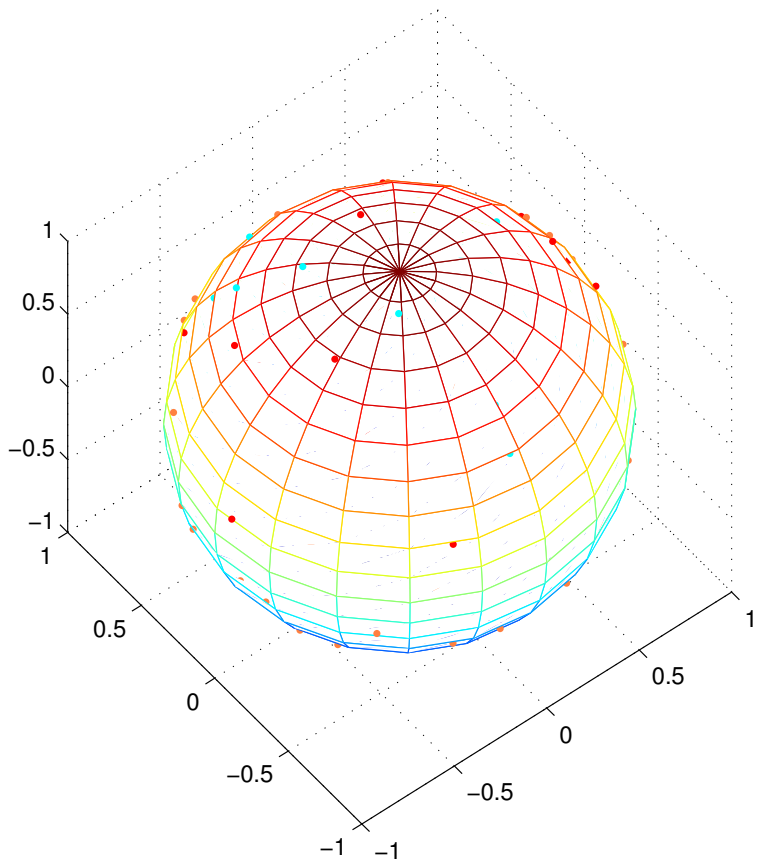
Scenario 2

$$f(X) = -\|(0,0,1) - X\|_2^{3/4}$$
$$\Pi \sim \text{Uniform}\left(S^2\right)$$
$$\sigma = 0.3$$



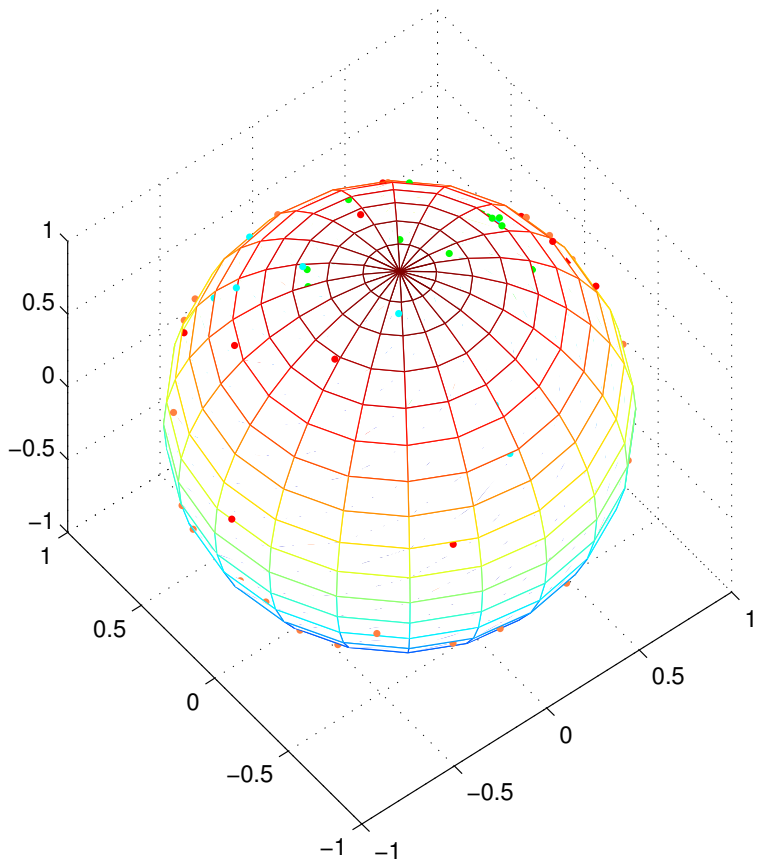
Scenario 2

$$f(X) = -\|(0,0,1) - X\|_2^{3/4}$$
$$\Pi \sim \text{Uniform}\left(S^2\right)$$
$$\sigma = 0.3$$



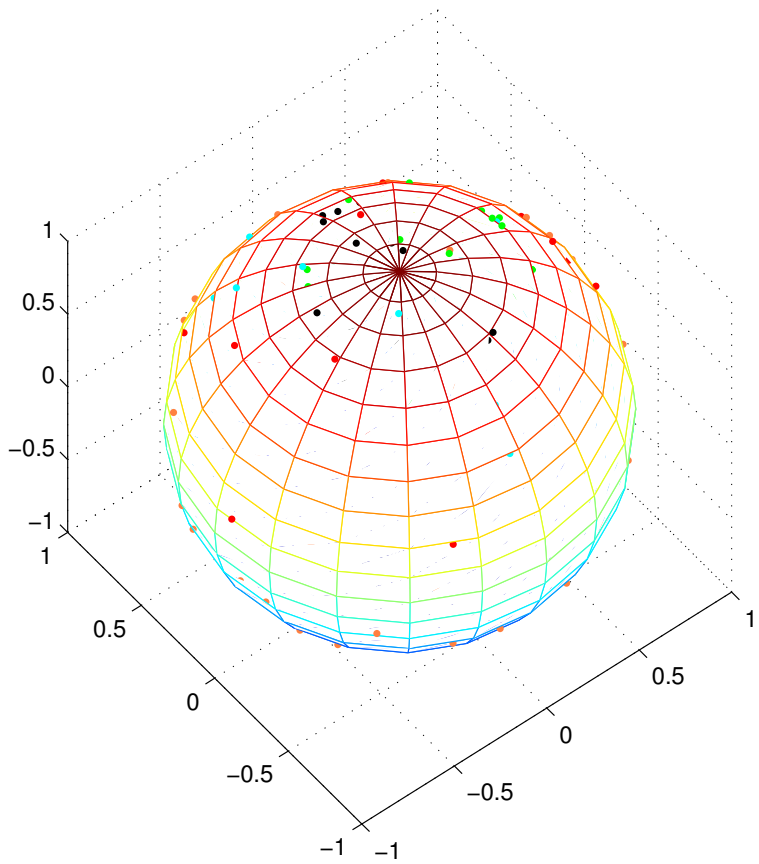
Scenario 2

$$f(X) = -\|(0,0,1) - X\|_2^{3/4}$$
$$\Pi \sim \text{Uniform}\left(S^2\right)$$
$$\sigma = 0.3$$



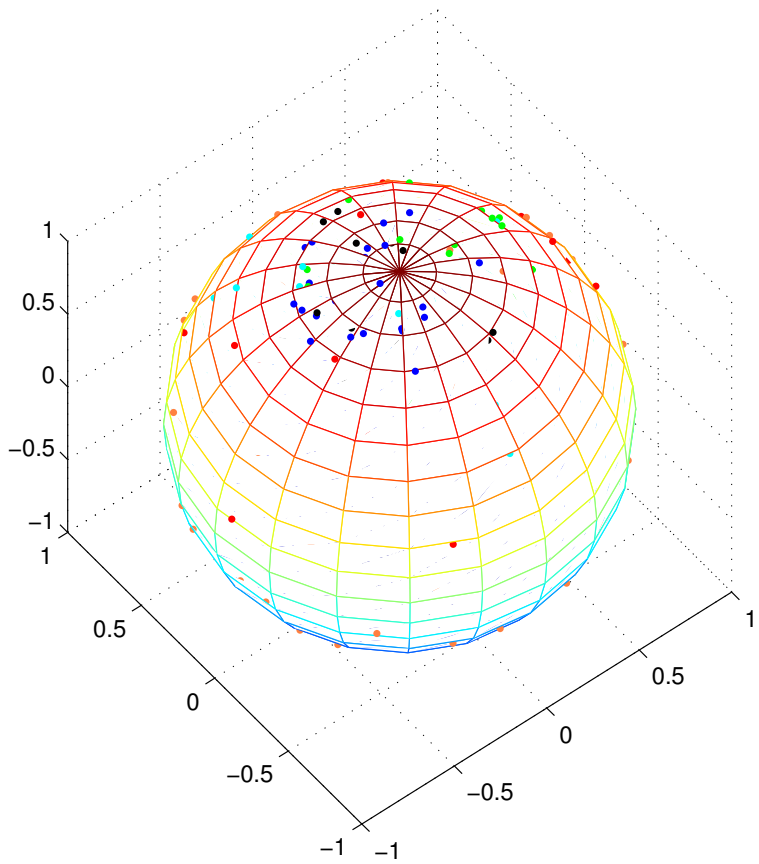
Scenario 2

$$f(X) = -\|(0,0,1) - X\|_2^{3/4}$$
$$\Pi \sim \text{Uniform}\left(S^2\right)$$
$$\sigma = 0.3$$



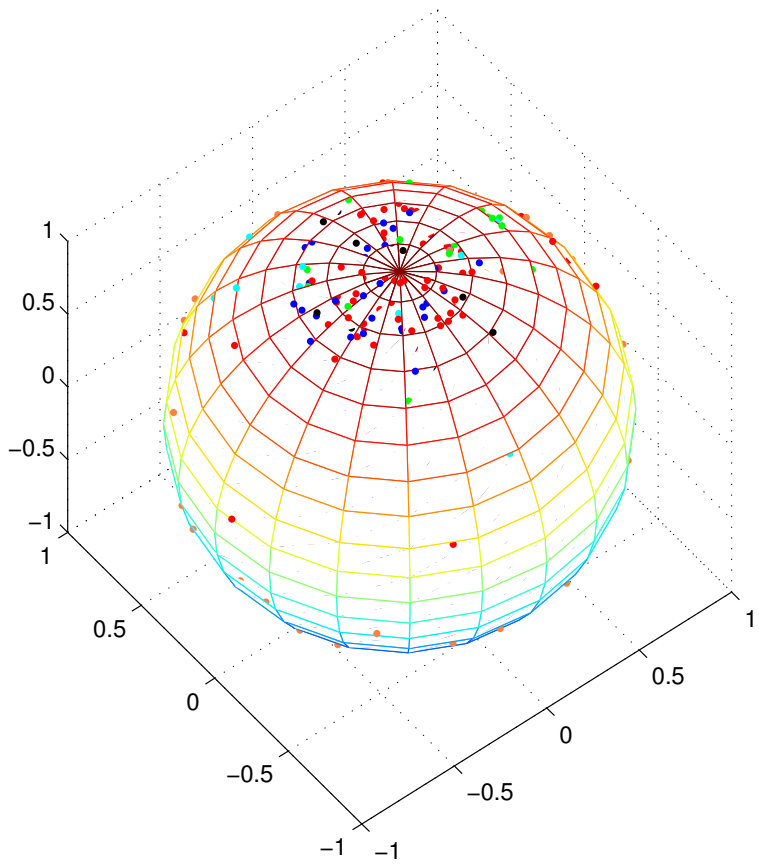
Scenario 2

$$f(X) = -\|(0,0,1) - X\|_2^{3/4}$$
$$\Pi \sim \text{Uniform}\left(S^2\right)$$
$$\sigma = 0.3$$



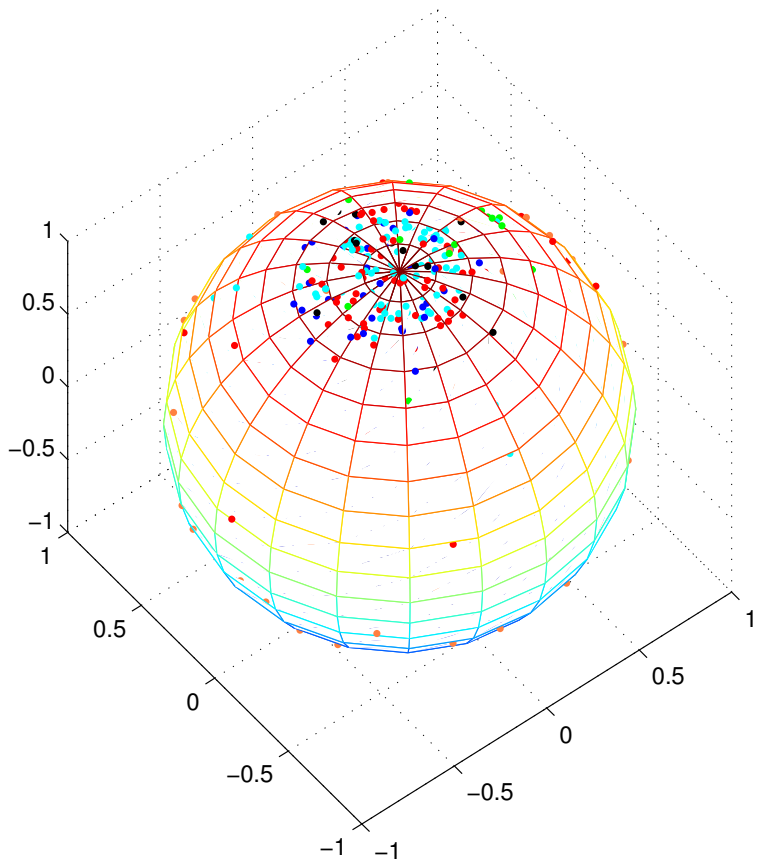
Scenario 2

$$f(X) = -\|(0,0,1) - X\|_2^{3/4}$$
$$\Pi \sim \text{Uniform}\left(S^2\right)$$
$$\sigma = 0.3$$



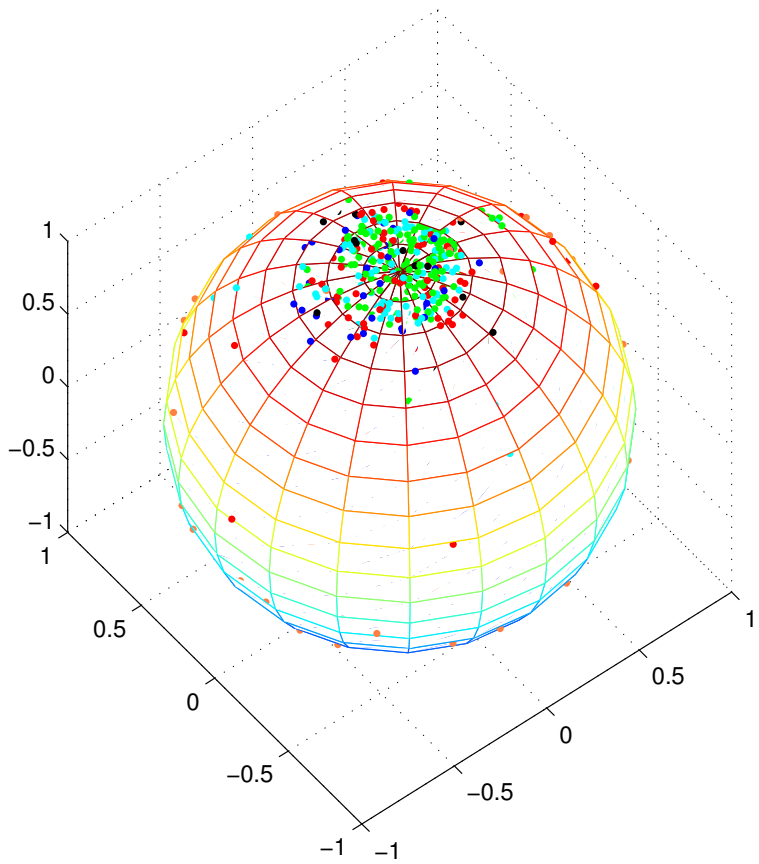
Scenario 2

$$f(X) = -\|(0,0,1) - X\|_2^{3/4}$$
$$\Pi \sim \text{Uniform}\left(S^2\right)$$
$$\sigma = 0.3$$



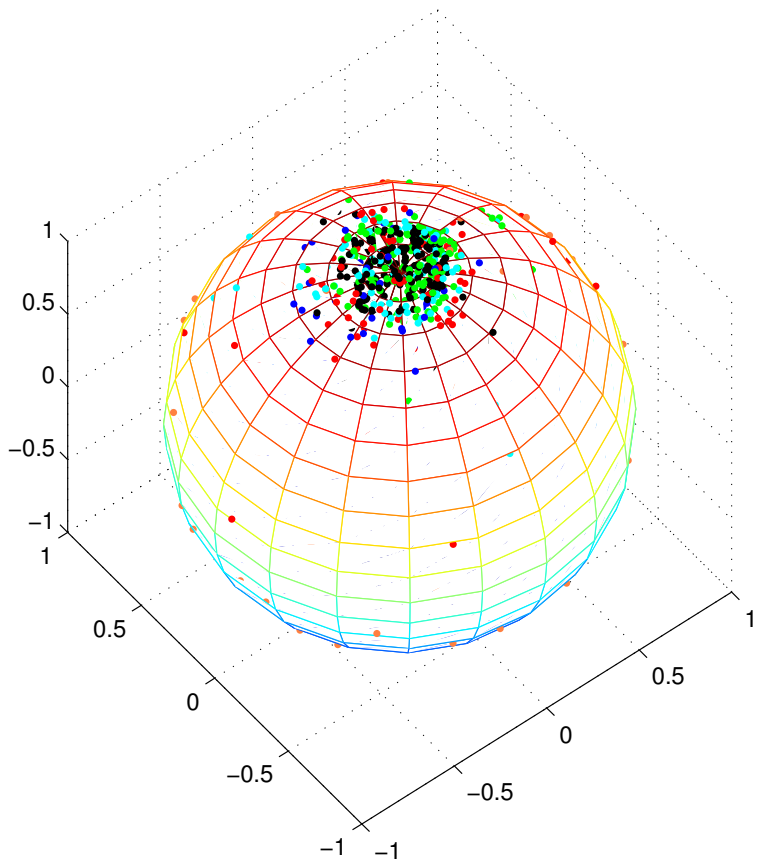
Scenario 2

$$f(X) = -\|(0,0,1) - X\|_2^{3/4}$$
$$\Pi \sim \text{Uniform}\left(S^2\right)$$
$$\sigma = 0.3$$



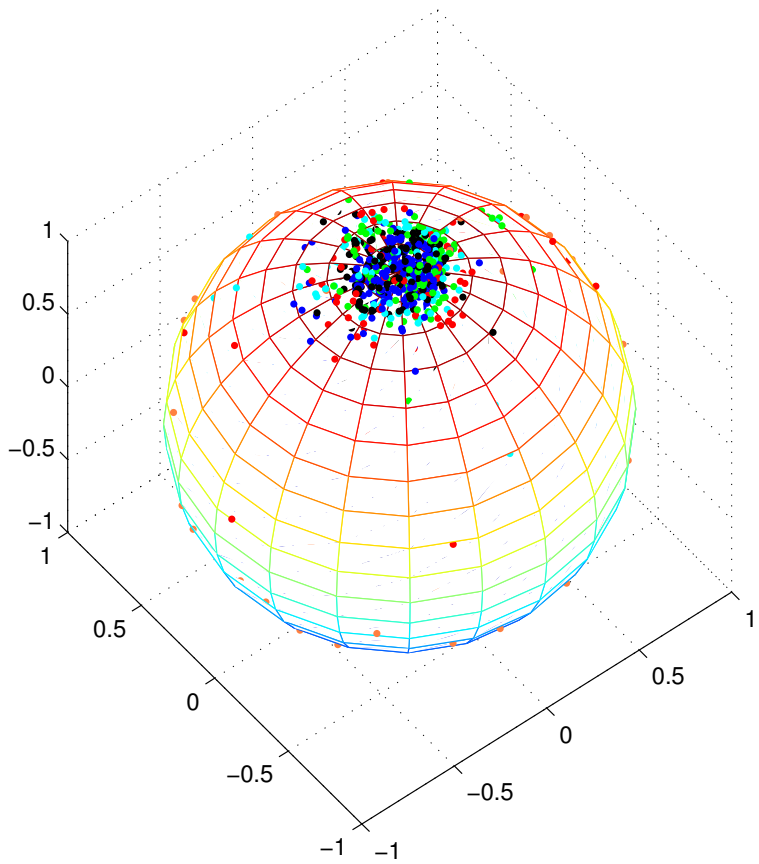
Scenario 2

$$f(X) = -\|(0,0,1) - X\|_2^{3/4}$$
$$\Pi \sim \text{Uniform}\left(S^2\right)$$
$$\sigma = 0.3$$



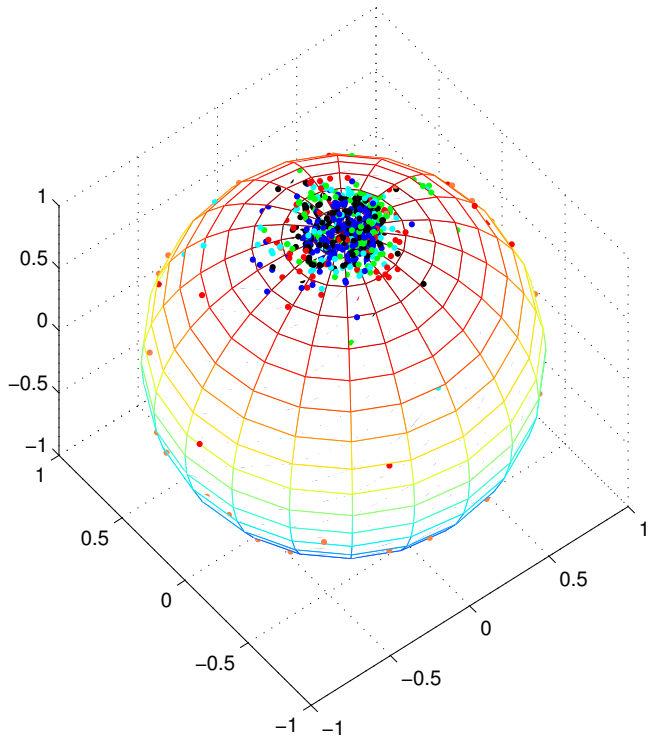
Scenario 2

$$f(X) = -\|(0,0,1) - X\|_2^{3/4}$$
$$\Pi \sim \text{Uniform}\left(S^2\right)$$
$$\sigma = 0.3$$



Scenario 2

$$f(X) = -\|(0,0,1) - X\|_2^{3/4}$$
$$\Pi \sim \text{Uniform}\left(S^2\right)$$
$$\sigma = 0.3$$



Estimated values:
max $= -0.2231$
# of function evaluations $= 638$
Total # of processed samples $= 15360$.

Thank you for your attention!