Solution Manifold and Its Statistical Applications

Yen-Chi Chen

Department of Statistics
University of Washington
◦ Supported by NSF DMS - 1810960 and DMS - 195278, NIH U01 - AG0169761

Solution manifolds

- A solution manifold is a manifold formed by the solutions of a system of equations (Rheinboldt 1988).

Solution manifolds

- A solution manifold is a manifold formed by the solutions of a system of equations (Rheinboldt 1988).
- Let $\Psi : \mathbb{R}^d \to \mathbb{R}^s$ be a system of s equations with d augments.
- The solution manifold generated by Ψ is

$$M = \{x : \Psi(x) = 0\}.$$

Solution manifolds

- A solution manifold is a manifold formed by the solutions of a system of equations (Rheinboldt 1988).
- Let $\Psi : \mathbb{R}^d \to \mathbb{R}^s$ be a system of s equations with d augments.
- The solution manifold generated by Ψ is

$$M = \{x : \Psi(x) = 0\}.$$

- Namely, the solution manifold is the solution set of a system of functions.
- We called Ψ the generator (function) of M.

Solution manifolds

- A solution manifold is a manifold formed by the solutions of a system of equations (Rheinboldt 1988).
- Let $\Psi : \mathbb{R}^d \to \mathbb{R}^s$ be a system of s equations with d augments.
- The solution manifold generated by Ψ is

$$M = \{x : \Psi(x) = 0\}.$$

- Namely, the solution manifold is the solution set of a system of functions.
- We called Ψ the generator (function) of M.
- Although the construct of a solution manifold seems to be abstract, it appears in many statistical problems.

Example: constrained likelihood

- Let $Y_1, \cdots, Y_n \sim N(\mu, \sigma^2)$, where μ and σ^2 are unknown parameters.

Example: constrained likelihood

- Let $Y_1, \cdots, Y_n \sim N(\mu, \sigma^2)$, where μ and σ^2 are unknown parameters.
- Suppose that we want to test the hypothesis

$$H_0 : P(-5 < Y_1 < 2) = 0.5.$$

Example: constrained likelihood

- Let $Y_1, \cdots, Y_n \sim N(\mu, \sigma^2)$, where μ and σ^2 are unknown parameters.
- Suppose that we want to test the hypothesis

$$H_0 : P(-5 < Y_1 < 2) = 0.5.$$

- There is one constraint ($s = 1$) and we have two parameters ($d = 2$).
- So the parameter space under H_0 forms a solution manifold.

Example: constrained likelihood

- Let $Y_1, \cdots, Y_n \sim N(\mu, \sigma^2)$, where μ and σ^2 are unknown parameters.
- Suppose that we want to test the hypothesis

$$H_0 : P(-5 < Y_1 < 2) = 0.5.$$

- There is one constraint ($s = 1$) and we have two parameters ($d = 2$).
- So the parameter space under H_0 forms a solution manifold.
- In this case,

$$\Psi(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-5}^{2} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy - 0.5.$$

Example: mixture models with moment constraints

- Let $Y_1, \cdots, Y_n \in \mathbb{R}$ be IID random variables from an unknown distribution.
- We fit a 2-Gaussian mixture model to the data; namely, the PDF can be written as

$$p(y) = \rho\phi(y; \mu_1, \sigma_2^2) + (1-\rho)\phi(y; \mu_2, \sigma_2^2),$$

where $\phi(y; \mu, \sigma^2)$ is the PDF of a normal distribution with mean μ variance σ^2.

Example: mixture models with moment constraints

- Let $Y_1, \cdots, Y_n \in \mathbb{R}$ be IID random variables from an unknown distribution.
- We fit a 2-Gaussian mixture model to the data; namely, the PDF can be written as

$$p(y) = \rho\phi(y; \mu_1, \sigma_2^2) + (1-\rho)\phi(y; \mu_2, \sigma_2^2),$$

  where $\phi(y; \mu, \sigma^2)$ is the PDF of a normal distribution with mean μ variance σ^2.
- There are a total of 5 parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

Example: mixture models with moment constraints

- Let $Y_1, \cdots, Y_n \in \mathbb{R}$ be IID random variables from an unknown distribution.
- We fit a 2-Gaussian mixture model to the data; namely, the PDF can be written as

$$p(y) = \rho\phi(y; \mu_1, \sigma_2^2) + (1-\rho)\phi(y; \mu_2, \sigma_2^2),$$

  where $\phi(y; \mu, \sigma^2)$ is the PDF of a normal distribution with mean μ variance σ^2.
- There are a total of 5 parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.
- Consider matching the first two moments to the data:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \rho\mu_1 + (1-\rho)\mu_2,$$

$$\frac{1}{n}\sum_{i=1}^{n} Y_i^2 = \rho(\mu_1^2 + \sigma_1^2) + (1-\rho)(\mu_2^2 + \sigma_2^2)$$

Example: geometric features

- Consider a nonparametric density estimation problem where $X_1, \cdots, X_n \sim p$, where p is the underlying unknown PDF.
- Many geometric features of p are solution manifolds.

Example: geometric features

- Consider a nonparametric density estimation problem where $X_1, \cdots, X_n \sim p$, where p is the underlying unknown PDF.
- Many geometric features of p are solution manifolds.
- The λ-level set (Polonik 1995, Walther 1997):

$$\{x : p(x) - \lambda = 0\}.$$

- The critical points:

$$\{x : \nabla p(x) = 0\}.$$

- The k-ridges (Genovese et al. 2014):

$$\{x : V_k(x) \nabla p(x) = 0, \lambda_{d-k} < 0\},$$

where $V_k(x)$ is the matrix of eigenvectors of the Hessian matrix corresponding to the $(d-k)$ smallest eigenvalues.

Solution manifolds

- In this talk, we will discuss both geometric and computational properties of solution manifolds.
- We will propose a gradient descent algorithm to compute the manifold.

Solution manifolds

- In this talk, we will discuss both geometric and computational properties of solution manifolds.
- We will propose a gradient descent algorithm to compute the manifold.
- Geometric properties:
  - **Smoothness**: how smooth the manifold is?
  - **Stability**: if we perturb the generator a bit, how much the manifold can change?

Solution manifolds

- In this talk, we will discuss both geometric and computational properties of solution manifolds.
- We will propose a gradient descent algorithm to compute the manifold.
- Geometric properties:
  - **Smoothness**: how smooth the manifold is?
  - **Stability**: if we perturb the generator a bit, how much the manifold can change?
- Computational properties:
  - **Gradient flow convergence**: when will the gradient flow converges to the manifold?
  - **Local manifold properties**: will the basin of attraction of a point on the manifold forms another manifold?
  - **Gradient descent algorithm convergence**: will the gradient descent converges? how fast it converges?

Assumptions

- Let the gradient and Hessian be
  $G_\Psi(x) = \nabla\Psi(x) \in \mathbb{R}^{s\times d}, \quad H_\Psi(x) = \nabla\nabla\Psi(x) \in \mathbb{R}^{s\times d\times d}.$

- Let the gradient and Hessian be
  $G_\Psi(x) = \nabla\Psi(x) \in \mathbb{R}^{s\times d}, \quad H_\Psi(x) = \nabla\nabla\Psi(x) \in \mathbb{R}^{s\times d\times d}.$
- Define

$$\|\Psi\|_{2,\infty}^* = \max\left\{\sup_x \|\Psi(x)\|_{\max}, \sup_x \|G_\Psi(x)\|_{\max}, \sup_x \|H_\Psi(x)\|_{\max}\right\}.$$

Assumptions

- Let the gradient and Hessian be
  $G_\Psi(x) = \nabla\Psi(x) \in \mathbb{R}^{s\times d}, \quad H_\Psi(x) = \nabla\nabla\Psi(x) \in \mathbb{R}^{s\times d\times d}.$
- Define

$$\|\Psi\|_{2,\infty}^{*} = \max\left\{\sup_x \|\Psi(x)\|_{\max}, \sup_x \|G_\Psi(x)\|_{\max}, \sup_x \|H_\Psi(x)\|_{\max}\right\}.$$

- For a set A, define $A \oplus r = \{x : d(x, A) \leq r\}$.

Assumptions

- Let the gradient and Hessian be
  $G_\Psi(x) = \nabla\Psi(x) \in \mathbb{R}^{s\times d}, \quad H_\Psi(x) = \nabla\nabla\Psi(x) \in \mathbb{R}^{s\times d\times d}.$
- Define

$$\|\Psi\|^*_{2,\infty} = \max\left\{\sup_x \|\Psi(x)\|_{\max}, \sup_x \|G_\Psi(x)\|_{\max}, \sup_x \|H_\Psi(x)\|_{\max}\right\}.$$

- For a set A, define $A \oplus r = \{x : d(x, A) \le r\}$.
- Consider the following assumptions:

(F1) Ψ is three-times bounded differentiable.

(F2) There exists $\lambda_0, \delta_0, c_0 > 0$ such that

1. $\lambda_{\min}(G_\Psi(x)G_\Psi(x)^T) \ge \lambda_0^2$ for all $x \in M \oplus \delta_0$.
2. $\|\Psi(x)\|_{\max} > c_0$ for all $x \notin M \oplus \delta_0$.

Smoothness of a solution manifold

Theorem (Smoothness theorem)

Assume (F1-2). Then

$$\mathsf{reach}(M) \geq \min\left\{\frac{\delta_0}{2}, \frac{\lambda_0}{\|\Psi\|_{2,\infty}^*}\right\}$$

- Reach (Federer 1959): the maximal distance that every point within this distance to M has a unique projection on M.
- This theorem links the smoothness of the generator Ψ into the smoothness of the solution manifold.

Stability of a solution manifold

- Let $\mathsf{Haus}(A, B) = \max\{\sup_{x\in A} d(x, B), \sup_{x\in B} d(x, A)\}$ be the Hausdorff distance between A and B.
- Let $\widetilde{\Psi} : \mathbb{R}^d \to \mathbb{R}^s$ be another generator function with at least bounded twice differentiable and $\widetilde{M}$ be its solution manifold.

- Let $\text{Haus}(A, B) = \max\{\sup_{x\in A} d(x, B), \sup_{x\in B} d(x, A)\}$ be the Hausdorff distance between A and B.
- Let $\widetilde{\Psi} : \mathbb{R}^d \to \mathbb{R}^s$ be another generator function with at least bounded twice differentiable and $\widetilde{M}$ be its solution manifold.

Theorem (Stability theorem)

Assume (F1-2) of Ψ. When $\|\Psi - \widetilde{\Psi}\|^*_{2,\infty}$ is sufficiently small,

- $\text{Haus}(M, \widetilde{M}) = O\left(\sup_x \|\Psi(x) - \widetilde{\Psi}(x)\|_{\max}\right)$.
- $\text{reach}(\widetilde{M}) \geq \min\left\{\frac{\delta_0}{2}, \frac{\lambda_0}{\|\Psi\|^*_{2,\infty}}\right\} + O\left(\|\Psi - \widetilde{\Psi}\|^*_{2,\infty}\right)$.

Consistency of a manifold estimator

- The stability theorem implies the consistency of a manifold estimator.

Consistency of a manifold estimator

- The stability theorem implies the consistency of a manifold estimator.
- Consider the 2-Gaussian mixture examples where the population solution manifold M is formed by

$$\mathbb{E}(Y_1) = \rho\mu_1 + (1-\rho)\mu_2, \quad \mathbb{E}(Y_1^2) = \rho(\mu_1^2 + \sigma_1^2) + (1-\rho)(\mu_2^2 + \sigma_2^2)$$

Consistency of a manifold estimator

- The stability theorem implies the consistency of a manifold estimator.
- Consider the 2-Gaussian mixture examples where the population solution manifold M is formed by

$$\mathbb{E}(Y_1) = \rho\mu_1 + (1-\rho)\mu_2, \quad \mathbb{E}(Y_1^2) = \rho(\mu_1^2 + \sigma_1^2) + (1-\rho)(\mu_2^2 + \sigma_2^2)$$

- The estimator of the solution manifold $\widehat{M}_n$ will be the one based on empirical moments:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \rho\mu_1 + (1-\rho)\mu_2,$$

$$\frac{1}{n}\sum_{i=1}^{n} Y_i^2 = \rho(\mu_1^2 + \sigma_1^2) + (1-\rho)(\mu_2^2 + \sigma_2^2)$$

Consistency of a manifold estimator

- The stability theorem implies the consistency of a manifold estimator.
- Consider the 2-Gaussian mixture examples where the population solution manifold M is formed by

$$\mathbb{E}(Y_1) = \rho\mu_1 + (1-\rho)\mu_2, \quad \mathbb{E}(Y_1^2) = \rho(\mu_1^2 + \sigma_1^2) + (1-\rho)(\mu_2^2 + \sigma_2^2)$$

- The estimator of the solution manifold $\widehat{M}_n$ will be the one based on empirical moments:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \rho\mu_1 + (1-\rho)\mu_2,$$
$$\frac{1}{n}\sum_{i=1}^{n} Y_i^2 = \rho(\mu_1^2 + \sigma_1^2) + (1-\rho)(\mu_2^2 + \sigma_2^2)$$

- The stability theorem shows that $\mathsf{Haus}(\widehat{M}_n, M) = O_P\left(\sqrt{\frac{1}{n}}\right)$.

Computing a solution manifold

- The above results characterize geometric properties of a solution manifold.
- But in practice, how do we numerically find the manifold?

Computing a solution manifold

- The above results characterize geometric properties of a solution manifold.
- But in practice, how do we numerically find the manifold?
- Here we propose a simple gradient descent algorithm to find the manifold (Boyd and Vandenberghe 2004).

Computing a solution manifold

- The above results characterize geometric properties of a solution manifold.
- But in practice, how do we numerically find the manifold?
- Here we propose a simple gradient descent algorithm to find the manifold (Boyd and Vandenberghe 2004).
- Let
$$f(x) = \Psi(x)^T \Psi(x) = \|\Psi(x)\|^2 \in \mathbb{R}.$$
- One may notice that
$$M = \{x : \Psi(x) = 0\} = \{x : f(x) = 0\}.$$

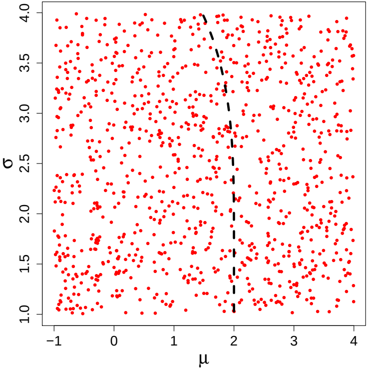
Computing a solution manifold

- The above results characterize geometric properties of a solution manifold.
- But in practice, how do we numerically find the manifold?
- Here we propose a simple gradient descent algorithm to find the manifold (Boyd and Vandenberghe 2004).
- Let
$$f(x) = \Psi(x)^T \Psi(x) = \|\Psi(x)\|^2 \in \mathbb{R}.$$
- One may notice that
$$M = \{x : \Psi(x) = 0\} = \{x : f(x) = 0\}.$$
- So we will find M by minimizing f.

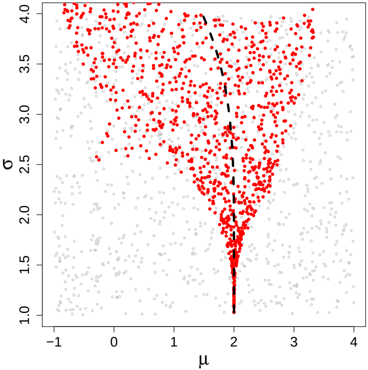
A gradient descent algorithm

1. Randomly choose an initial point $x_0 \sim Q$, where Q is a distribution over the region of interest $\mathbb{K}$.
2. Iterates
$$x_{t+1} \leftarrow x_t - \gamma \nabla f(x_t)$$
until convergence. Let x_∞ be the convergent point.
3. If $\Psi(x_\infty) = 0$ (or sufficiently small), we keep x_∞; otherwise, discard x_∞.
4. Repeat the above procedure until we obtain enough points for approximating M.

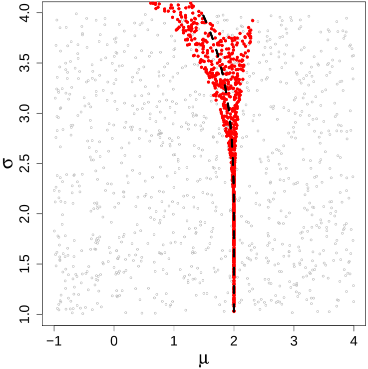
Gradient descent: illustration



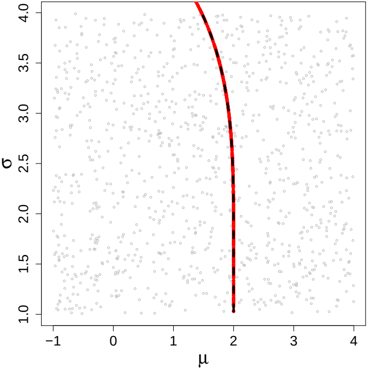
Gradient descent: illustration



Gradient descent: illustration



Gradient descent: illustration



Gradient flow

- To study how the gradient descent algorithm works, we first analyze the (continuous-time) gradient flow $\pi : \mathbb{R} \to \mathbb{R}^d$

$$\pi_x(0) = x, \quad \pi_x'(t) = -\nabla f(\pi_x(t)).$$

Gradient flow

- To study how the gradient descent algorithm works, we first analyze the (continuous-time) gradient flow $\pi : \mathbb{R} \to \mathbb{R}^d$

$$\pi_x(0) = x, \quad \pi_x'(t) = -\nabla f(\pi_x(t)).$$

- $\pi_x(\infty) = \lim_{t\to\infty} \pi_x(t)$ is called the destination of π_x.

Gradient flow

- To study how the gradient descent algorithm works, we first analyze the (continuous-time) gradient flow $\pi : \mathbb{R} \to \mathbb{R}^d$

$$\pi_x(0) = x, \quad \pi_x'(t) = -\nabla f(\pi_x(t)).$$

- $\pi_x(\infty) = \lim_{t\to\infty} \pi_x(t)$ is called the destination of π_x.
- Also, let $v_x(t) = \frac{\pi_x'(t)}{\|\pi_x'(t)\|}$ be the directional vector at time t and $v_x(\infty) = \lim_{t\to\infty} v_x(t)$.

Consistency of the gradient flow

Theorem (Gradient flow convergence)

Assume (F1-2) and let

$$\delta_c = \min\left\{\frac{\delta_0}{2}, \frac{1}{8d}\frac{\lambda_0^2}{\|\Psi\|_{2,\infty}^*\|\Psi\|_{3,\infty}^*}\right\}.$$

Then

- **Convergence radius.** *If $x \in M \oplus \delta_c$, $\pi_x(\infty) \in M$.*
- **Terminal flow orientation.** *If $\pi_x(\infty) \in M$, then $v_x(\infty) \perp M$ at $\pi_x(\infty)$.*

- Namely, if the initial point is within δ_c distance to M, the gradient flow converges to M.

Local stable manifold theorem

- For a point $z \in M$, its basin of attraction is

$$A(z) = \{x : \pi_x(\infty) = z\}.$$

- Namely, $A(z)$ is the collection of points converging to z by the gradient flow.

Local stable manifold theorem

- For a point $z \in M$, its basin of attraction is

$$A(z) = \{x : \pi_x(\infty) = z\}.$$

- Namely, $A(z)$ is the collection of points converging to z by the gradient flow.
- Interestingly, $A(z)$ forms another manifold, known as the local stable manifold of a gradient flow (Perko 2001).

Local stable manifold theorem

- For a point $z \in M$, its basin of attraction is

$$A(z) = \{x : \pi_x(\infty) = z\}.$$

- Namely, $A(z)$ is the collection of points converging to z by the gradient flow.
- Interestingly, $A(z)$ forms another manifold, known as the local stable manifold of a gradient flow (Perko 2001).

Theorem (Local stable manifold theorem)

Assume (F1-2). Then $A(z)$ forms an s-dimensional manifold for each $z \in M$.

Implication on manifold data

- Here is an interesting implication.
- If we initialize from a regular PDF q over $\mathbb{R}^d$, the convergent points forms a distribution Q_π over M such that Q_π has an $(d - s)$-dimensional Hausdorff density (Preiss 1987).

Implication on manifold data

- Here is an interesting implication.
- If we initialize from a regular PDF q over $\mathbb{R}^d$, the convergent points forms a distribution Q_π over M such that Q_π has an $(d-s)$-dimensional Hausdorff density (Preiss 1987).
- Specifically, suppose we have initial points $x_1, \cdots, x_n \sim q$ and let $z_1, \cdots, z_n$ be the corresponding points on the manifold M by the gradient flow.

Implication on manifold data

- Here is an interesting implication.
- If we initialize from a regular PDF q over $\mathbb{R}^d$, the convergent points forms a distribution Q_π over M such that Q_π has an $(d - s)$-dimensional Hausdorff density (Preiss 1987).
- Specifically, suppose we have initial points $x_1, \cdots, x_n \sim q$ and let $z_1, \cdots, z_n$ be the corresponding points on the manifold M by the gradient flow.
- Then $z_1, \cdots, z_n$ can be viewed as IID from a density on M.

Implication on manifold data

- Here is an interesting implication.
- If we initialize from a regular PDF q over $\mathbb{R}^d$, the convergent points forms a distribution Q_π over M such that Q_π has an $(d - s)$-dimensional Hausdorff density (Preiss 1987).
- Specifically, suppose we have initial points $x_1, \cdots, x_n \sim q$ and let $z_1, \cdots, z_n$ be the corresponding points on the manifold M by the gradient flow.
- Then $z_1, \cdots, z_n$ can be viewed as IID from a density on M.
- This becomes a scenario that IID observations on a manifold is a reasonable model.

Theory of gradient descent algorithm

- In reality, we use a discrete time gradient descent algorithm; namely, we use the discrete update:

$$x_{t+1} = x_t - \gamma \nabla f(x_t)$$

and $\gamma > 0$ is the step size.

Theory of gradient descent algorithm

- In reality, we use a discrete time gradient descent algorithm; namely, we use the discrete update:

$$x_{t+1} = x_t - \gamma \nabla f(x_t)$$

  and $\gamma > 0$ is the step size.
- When $\gamma \approx 0$, the algorithm behaves just like the gradient flow.

Theory of gradient descent algorithm

- In reality, we use a discrete time gradient descent algorithm; namely, we use the discrete update:

$$x_{t+1} = x_t - \gamma \nabla f(x_t)$$

  and $\gamma > 0$ is the step size.
- When $\gamma \approx 0$, the algorithm behaves just like the gradient flow.
- We proved that when γ is sufficiently small and x_0 is properly initialized,

$$f(x_K) \leq f(x_0) \cdot \left(1 - \gamma \frac{\lambda_0^4}{\|\Psi\|_{2,\infty}^*}\right)^K$$

$$d(x_K, M) \leq d(x_0, M) \cdot \left(1 - \gamma \lambda_0^2\right)^{K/2}.$$

  for each $K = 1, 2, 3, \cdots$.

Theory of gradient descent algorithm

- In reality, we use a discrete time gradient descent algorithm; namely, we use the discrete update:

$$x_{t+1} = x_t - \gamma \nabla f(x_t)$$

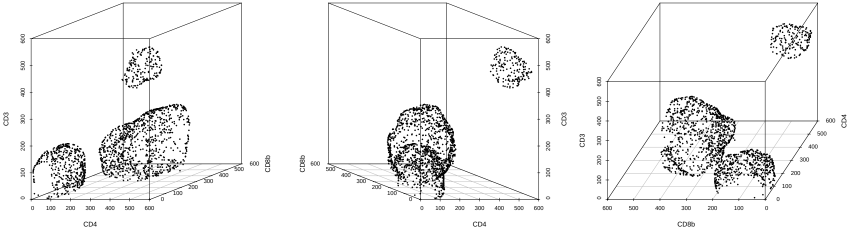
  and $\gamma > 0$ is the step size.
- When $\gamma \approx 0$, the algorithm behaves just like the gradient flow.
- We proved that when γ is sufficiently small and x_0 is properly initialized,

$$f(x_K) \leq f(x_0) \cdot \left(1 - \gamma \frac{\lambda_0^4}{\|\Psi\|_{2,\infty}^*}\right)^K$$

$$d(x_K, M) \leq d(x_0, M) \cdot \left(1 - \gamma \lambda_0^2\right)^{K/2}.$$

  for each $K = 1, 2, 3, \cdots$.
- An interesting fact: f is a non-convex function so we are using gradient descent on a non-convex function.

A 2D manifold example



- This is the density level sets in a $3D$ data (GvHD data in R); the level sets form 2-dimensional manifolds.
- The three panels are three different view of the level sets.

Discussion: assumptions

- One may notice that all five theorems rely on the same set of assumptions:
  - (F1) Ψ is three-times bounded differentiable.
  - (F2) There exists $\lambda_0, \delta_0, c_0 > 0$ such that
    1. $\lambda_{\min}(G\Psi(x)G\Psi(x)^T) \geq \lambda_0$ for all $x \in M \oplus \delta_0$.
    2. $\|\Psi(x)\|_{\max} > c_0$ for all $x \notin M \oplus \delta_0$.

Discussion: assumptions

- One may notice that all five theorems rely on the same set of assumptions:
  - (F1) Ψ is three-times bounded differentiable.
  - (F2) There exists $\lambda_0, \delta_0, c_0 > 0$ such that
    1. $\lambda_{\min}(G\Psi(x)G\Psi(x)^T) \geq \lambda_0$ for all $x \in M \oplus \delta_0$.
    2. $\|\Psi(x)\|_{\max} > c_0$ for all $x \notin M \oplus \delta_0$.
- This shows that the smoothness, stability, gradient flow, and gradient descent algorithm are all implicitly related.

Discussion: assumptions

- One may notice that all five theorems rely on the same set of assumptions:
  - (F1) Ψ is three-times bounded differentiable.
  - (F2) There exists $\lambda_0, \delta_0, c_0 > 0$ such that
    1. $\lambda_{\min}(G_\Psi(x)G_\Psi(x)^T) \geq \lambda_0$ for all $x \in M \oplus \delta_0$.
    2. $\|\Psi(x)\|_{\max} > c_0$ for all $x \notin M \oplus \delta_0$.
- This shows that the smoothness, stability, gradient flow, and gradient descent algorithm are all implicitly related.
- In fact, this is a generic result that other M-estimator also share but somehow we did not emphasize this in statistics.

- One may notice that all five theorems rely on the same set of assumptions:
  - (F1) Ψ is three-times bounded differentiable.
  - (F2) There exists $\lambda_0, \delta_0, c_0 > 0$ such that
    1. $\lambda_{\min}(G_\Psi(x)G_\Psi(x)^T) \geq \lambda_0$ for all $x \in M \oplus \delta_0$.
    2. $\|\Psi(x)\|_{\max} > c_0$ for all $x \notin M \oplus \delta_0$.
- This shows that the smoothness, stability, gradient flow, and gradient descent algorithm are all implicitly related.
- In fact, this is a generic result that other M-estimator also share but somehow we did not emphasize this in statistics.
- Note: for some theorems, these two assumptions are often stronger than what we actually need but unifying them give us some new insights.

Discussion: connections to other fields

- **Econometrics.** The generalized method of moments (Hansen 1982) is tightly connected to solution manifolds. In particular, they are often using the minimizer of a function f as a numerical method for finding a solution.

Discussion: connections to other fields

- **Econometrics.** The generalized method of moments (Hansen 1982) is tightly connected to solution manifolds. In particular, they are often using the minimizer of a function f as a numerical method for finding a solution.
- **Dynamical system.** The local stable manifold theorem is from dynamical system literature (Perko 2001). Here we present a new use of this theorem on data analysis.

Discussion: connections to other fields

- **Econometrics.** The generalized method of moments (Hansen 1982) is tightly connected to solution manifolds. In particular, they are often using the minimizer of a function f as a numerical method for finding a solution.
- **Dynamical system.** The local stable manifold theorem is from dynamical system literature (Perko 2001). Here we present a new use of this theorem on data analysis.
- **Computational geometry.** Numerically computing a manifold is a classical problem in computational geometry (Dey 2006). Here we present a set of new procedures for this purposes and analyze the underlying algorithmic properties.

Discussion: connections to other fields

- **Econometrics.** The generalized method of moments (Hansen 1982) is tightly connected to solution manifolds. In particular, they are often using the minimizer of a function f as a numerical method for finding a solution.
- **Dynamical system.** The local stable manifold theorem is from dynamical system literature (Perko 2001). Here we present a new use of this theorem on data analysis.
- **Computational geometry.** Numerically computing a manifold is a classical problem in computational geometry (Dey 2006). Here we present a set of new procedures for this purposes and analyze the underlying algorithmic properties.
- **Optimization.** We show that for a particular family of non-convex function f, the gradient descent may still converge quickly. This may reveal a new class of non-convex problem that is easy to solve.

Thank You!

More details can be found in `https://arxiv.org/abs/2002.05297`.

References

1. Y.-C. Chen. Solution manifold and Its Statistical Applications. arXiv preprint arXiv:2002.05297 (2020).
2. W. C. Rheinboldt. On the computation of multi-dimensional solution manifolds of parametrized equations. Numerische Mathematik, 1988.
3. G. Walther. Granulometric smoothing. The Annals of Statistics, 25(6):2273-2299, 1997.
4. W. Polonik. Measuring mass concentrations and estimating density contour clusters-an excess mass approach. The Annals of Statistics, 23(3), pp.855-881.
5. C. R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Nonparametric ridge estimation. The Annals of Statistics, 42(4):1511-1545, 2014.
6. H. Federer. Curvature measures. Transactions of the American Mathematical Society, 93 (3):418-491, 1959.
7. S. Boyd and L. Vandenberghe. Convex optimization. Cambridge university press, 2004.
8. L. Perko. Differential equations and dynamical systems, volume 7. Springer Science & Business Media, 2001.
9. D. Preiss. Geometry of measures in r n: distribution, rectifiability, and densities. Annals of Mathematics, 125(3):537-643, 1987.
10. L. P. Hansen. Large sample properties of generalized method of moments estimators. Econometrica: Journal of the Econometric Society, pages 1029-1054, 1982.
11. T. K. Dey. Curve and surface reconstruction: algorithms with mathematical analysis, volume 23. Cambridge University Press, 2006.

Reach of a manifold

- By the implicit function theorem, if the rank of the matrix $\nabla\Psi(x)$ is s, the same as the number of equations, then M is an $(d - s)$ dimensional manifold.
- But this does not tell us anything about the smoothness of M

- By the implicit function theorem, if the rank of the matrix $\nabla\Psi(x)$ is s, the same as the number of equations, then M is an $(d-s)$ dimensional manifold.
- But this does not tell us anything about the smoothness of M
- To quantify the smoothness, we use the concept of *reach*:

$$\mathsf{reach}(M) = \sup\{r : x \text{ has a unique projection onto } M \text{ for all } d(x,M) \le r\},$$

where $d(x,M) = \inf_{y\in M} \|x-y\|$ is the projection distance from x to M.

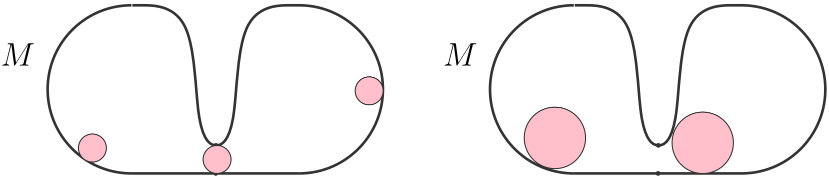
Reach of a manifold

- By the implicit function theorem, if the rank of the matrix $\nabla\Psi(x)$ is s, the same as the number of equations, then M is an $(d-s)$ dimensional manifold.
- But this does not tell us anything about the smoothness of M
- To quantify the smoothness, we use the concept of *reach*:

  $$\text{reach}(M) = \sup\{r : x \text{ has a unique projection onto } M \text{ for all } d(x,M) \leq r\},$$

  where $d(x,M) = \inf_{y\in M} \|x-y\|$ is the projection distance from x to M.
- A simple way to think of a reach is via its ball-rolling property.

Example: reach



- If r is less than the reach, then a ball with radius r can roll freely around the manifold (left panel).
- If r is larger than the reach, then a ball with radius r cannot roll freely around the manifold (right panel).

Theory of gradient descent algorithm - 1

Theorem (Convergence of gradient decent algorithm)

Assume (F1-2) and let δ_c *be the same as the theorem of gradient flow. Suppose that the step size satisfies*

$$\gamma < \min\left\{ \frac{1}{\|\Psi\|_{2,\infty}^*}, \frac{\|\Psi\|_{2,\infty}^*}{4\lambda_0^2}, \delta_c \right\}$$

and $d(x_0, M) \leq \delta_c$. *Then for each* $T = 1, 2, 3, \cdots$

$$f(x_T) \leq f(x_0) \cdot \left(1 - \gamma \frac{\lambda_0^4}{\|\Psi\|_{2,\infty}^*}\right)^T$$

$$d(x_T, M) \leq d(x_0, M) \cdot \left(1 - \gamma \lambda_0^2\right)^{T/2}.$$

Theory of gradient descent algorithm - 2

$$f(x_T) \leq f(x_0) \cdot \left(1 - \gamma \frac{\lambda_0^4}{\|\Psi\|_{2,\infty}^*}\right)^T$$

$$d(x_T, M) \leq d(x_0, M) \cdot \left(1 - \gamma \lambda_0^2\right)^{T/2}$$

- An equivalent statement is that the algorithm takes $O(\log(1/\epsilon))$ to converges to ϵ-error to the minimum.

Theory of gradient descent algorithm - 2

$$f(x_T) \leq f(x_0) \cdot \left(1 - \gamma \frac{\lambda_0^4}{\|\Psi\|_{2,\infty}^*}\right)^T$$

$$d(x_T, M) \leq d(x_0, M) \cdot \left(1 - \gamma \lambda_0^2\right)^{T/2}$$

- An equivalent statement is that the algorithm takes $O(\log(1/\epsilon))$ to converges to ϵ-error to the minimum.
- The above convergence is also known as the linear convergence, a common result in convex optimization.

Theory of gradient descent algorithm - 2

$$f(x_T) \leq f(x_0) \cdot \left(1 - \gamma \frac{\lambda_0^4}{\|\Psi\|_{2,\infty}^*}\right)^T$$
$$d(x_T, M) \leq d(x_0, M) \cdot \left(1 - \gamma \lambda_0^2\right)^{T/2}$$

- An equivalent statement is that the algorithm takes $O(\log(1/\epsilon))$ to converges to ϵ-error to the minimum.
- The above convergence is also known as the linear convergence, a common result in convex optimization.
- An interesting fact: f is a non-convex function so we are using gradient descent on a non-convex function. But we still obtain a similar result to a convex problem.

Extension 1: manifold-constraint maximization

- In likelihood inference, finding the manifold is often not the final goal.
- What we need is the MLE on the manifold.
- Here we propose an alternating algorithm consisting of two major steps: ascent of likelihood and descent to the manifold.

Manifold-constraint maximizing algorithm

1. Randomly choose an initial point $\theta_0^{(0)} = \theta_\infty^{(0)} \in \Theta$.
2. For $m = 1, 2, \cdots$, do step 3-6:
3. **Ascent of likelihood.** Update

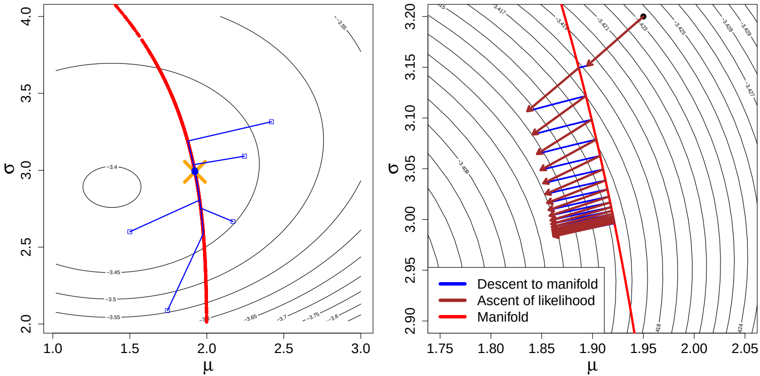
$$\theta_0^{(m)} = \theta_\infty^{(m-1)} + \alpha \nabla \ell(\theta_\infty^{(m-1)} | X_1, \cdots, X_n),$$

   where $\alpha > 0$ is the step size of the gradient ascent over likelihood function and $\ell(\theta | X_1, \cdots, X_n)$ is the log-likelihood function.
4. **Descent to manifold.** For each $t = 0, 1, 2, \cdots$ iterates

$$\theta_{t+1}^{(m)} \leftarrow \theta_t^{(m)} - \gamma \nabla f(\theta_t^{(m)})$$

   until convergence. Let $\theta_\infty^{(m)}$ be the convergent point.
5. If $\Psi(\theta_\infty^{(m)}) = 0$ (or sufficiently small), we keep $\theta_\infty^{(m)}$; otherwise, discard $\theta_\infty^{(m)}$ and return to step 1.
6. If $\nabla \ell(\theta_\infty^{(m)} | X_1, \cdots, X_n)$ belongs to the row space of $\nabla \Psi(\theta_\infty^{(m)})$, we stop and output $\theta_\infty^{(m)}$.

Illustration: manifold-constraint maximization



Extension 2: approximating a posterior on a manifold

- Suppose that we place a prior distribution $\pi(\theta)$ over a solution manifold M, i.e.,

$$\pi(\theta) = 0 \text{ if } \theta \notin M.$$

Extension 2: approximating a posterior on a manifold

- Suppose that we place a prior distribution $\pi(\theta)$ over a solution manifold M, i.e.,

$$\pi(\theta) = 0 \text{ if } \theta \notin M.$$

- And then we observe data $Y_1, \cdots, Y_n$ so we will update the prior to be the posterior distribution $\pi(\theta | Y_1, \cdots, Y_n)$.

Extension 2: approximating a posterior on a manifold

- Suppose that we place a prior distribution $\pi(\theta)$ over a solution manifold M, i.e.,
$$\pi(\theta) = 0 \text{ if } \theta \notin M.$$
- And then we observe data $Y_1, \cdots, Y_n$ so we will update the prior to be the posterior distribution $\pi(\theta|Y_1, \cdots, Y_n)$.
- One may be wondering how do we represent the posterior distribution in this case.

Extension 2: approximating a posterior on a manifold

- Suppose that we place a prior distribution $\pi(\theta)$ over a solution manifold M, i.e.,

$$\pi(\theta) = 0 \text{ if } \theta \notin M.$$

- And then we observe data $Y_1, \cdots, Y_n$ so we will update the prior to be the posterior distribution $\pi(\theta | Y_1, \cdots, Y_n)$.
- One may be wondering how do we represent the posterior distribution in this case.
- Here we propose a simple approach to approximate the posterior distribution.

Approximated manifold posterior algorithm

1. Generate many points $Z_1, \cdots, Z_N \in M$ by the gradient descent.
2. Estimate a density score of Z_i using

$$\widehat{\rho}_{i,N} = \frac{1}{N} \sum_{j=1}^{N} K\left(\frac{\|Z_i - Z_j\|}{h}\right),$$

   where $h > 0$ is a tuning parameter and K is a smooth function such as a Gaussian.
3. Compute the posterior density score of Z_i as

$$\widehat{\omega}_{i,N} = \frac{1}{\widehat{\rho}_{i,N}} \cdot \widehat{\pi}_{i,N}, \quad \widehat{\pi}_{i,N} = \pi(Z_i) \cdot \prod_{j=1}^{n} p(X_j | Z_i),$$

4. Return: Weighted point clouds $(Z_1, \widehat{\omega}_{i,N}), \cdots, (Z_N, \widehat{\omega}_{N,N})$.

Illustration: approximated manifold posterior