Numerical Methods for Wasserstein Natural Gradient Descent in Inverse Problems

NYU SURE Program Presentation
Wanzhou Lei Oct 2021

Mentor: Professor Yunan Yang

Levon Nurbekyan, Dept of Math, UCLA
Yunan Yang, Dept of Math, Cornell University
Wanzhou Lei, New York University

Background and Motivation

Consider the following optimization problem:

$$\inf_{\theta} f(\rho(\theta)) = \inf_{\theta} d(\rho(\theta), \rho^*)$$

ρ : Synthetic based on θ
ρ*: Observed Data

Standard Gradient Descent:

$$\theta^{l+1} = \theta^l - \tau \partial_\theta f$$

τ is step size

$$\theta^{l+1} = \underset{\theta}{\operatorname{argmin}} \{ f(\rho(\theta)) + \frac{||\theta - \theta^l||^2}{2\tau} \}$$

Proximal Operator

2 Cons:

1. Depends on parameterizations.
2. Slow convergence. (first-order method)

Background and Motivation

Natural Gradient Descent

$$\theta_{std}^{l+1} = \underset{\theta}{\operatorname{argmin}}\{f(\rho(\theta)) + \frac{||\theta - \theta^l||^2}{2\tau}\}$$

$$\theta_{nat}^{l+1} = \underset{\theta}{\operatorname{argmin}}\{f(\rho(\theta)) + \frac{d_\rho(\rho(\theta), \rho(\theta^l))}{2\tau}\}$$

Where τ is the step size.

d_ρ is the metric in ρ - space. (data space)

Mathematical Formulations of Natural Gradient Descent

ρ lives in (M, g). If ρ: θ → ρ(θ) is smooth, there exists tangent vectors:

$$\{\partial^g_{\theta_1}\rho(\theta), \partial^g_{\theta_2}\rho(\theta), ..., \partial^g_{\theta_p}\rho(\theta)\} \subset T_\rho M$$

If f is a smooth loss function, then for all smooth curves t →ρ (t) in ρ - space, by chain rule, we have:

$$\frac{df(\rho(t))}{dt} = \left\langle \partial^g_\rho f(\rho(t)), \partial^g_t \rho(t) \right\rangle_{g(\rho(t))}$$

Mathematical Formulations of Natural Gradient Descent

If the derivative w.r.t t of the curve in θ - space is:

$$\frac{d\theta}{dt} = \eta = (\eta_1, \dots, \eta_p)^\top \in \mathbb{R}^p$$

$$\Rightarrow \quad \partial_t^g \rho(\theta) = \partial_\theta^g \rho(\theta) \frac{d\theta}{dt} = \eta_1 \partial_{\theta_1}^g \rho + \cdots + \eta_p \partial_{\theta_p}^g \rho$$

$$\Rightarrow \quad \frac{df(\rho(\theta))}{dt} = \left\langle \partial_\rho^g f, \partial_t^g \rho(\theta) \right\rangle_{g(\rho(\theta))} = \left\langle \partial_\rho^g f, \sum_{i=1}^{p} \eta_i \partial_{\theta_i}^g \rho \right\rangle_{g(\rho(\theta))}$$

Mathematical Formulations of Natural Gradient Descent

$$\frac{df(\rho(\theta))}{dt} = \left\langle \partial_\rho^g f, \partial_t^g \rho(\theta) \right\rangle_{g(\rho(\theta))} = \left\langle \partial_\rho^g f, \sum_{i=1}^{p} \eta_i \partial_{\theta_i}^g \rho \right\rangle_{g(\rho(\theta))}$$

$$\sum_{i=1}^{p} \eta_i \partial_{\theta_i}^g \rho = -\partial_\rho^g f \quad \Rightarrow \quad \eta^{nat} = \operatorname*{argmin}_{\eta} \left\| \partial_\rho^g f + \sum_{i=1}^{p} \eta_i \partial_{\theta_i}^g \rho \right\|_{g(\rho(\theta))}^2$$

Method #1: Solving least square linear system

Least square solution of:

$$(\partial_\theta^g \rho)\eta = -\partial_\rho^g f$$

Mathematical Formulations of Natural Gradient Descent

Method #2: Compute the Information Matrix

$$(\partial_\theta^g \rho)\eta = -\partial_\rho^g f$$

$$\Rightarrow \quad (\partial_\theta^g \rho^T \partial_\theta^g \rho)\eta^{nat} = -\partial_\theta^g \rho^T \partial_\rho^g f = -\partial_\theta f = \eta^{std}$$

(Normal Equation)

$$\Rightarrow \quad \eta^{nat} = (\partial_\theta^g \rho^T \partial_\theta^g \rho)^{-1} \eta^{std} = G_g(\theta)^{-1} \eta^{std}$$

Define Information Matrix:

$$(G_g(\theta))_{ij} = \left\langle \partial_{\theta_i}^g \rho, \partial_{\theta_j}^g \rho \right\rangle_{g(\rho(\theta))}, \quad i, j = 1, \ldots, p$$

More Common way. Good when p is small.

Mathematical Formulations of Natural Gradient Descent

Summary of Method #1 and Method #2

#1

$$\eta^{nat} = argmin_{\eta} \left\| \partial_{\rho}^{g} f + \sum_{i=1}^{p} \eta_i \partial_{\theta_i}^{g} \rho \right\|_{g(\rho(\theta))}^{2}.$$

#2

$$\eta^{nat} = (\partial_{\theta}^{g} \rho^{T} \partial_{\theta}^{g} \rho)^{-1} \eta^{std} = G_g(\theta)^{-1} \eta^{std}$$

1. G may not be invertible.
2. G is dense.
3. G's condition number is large.

Next we consider:

$$(M, g) = \begin{cases} L^2 & \text{metric} \\ W^2 & \text{metric} \end{cases}$$

L^2 Natural Gradient Descent Formulation:

Metric space: $(M, g) = (L^2(\mathbb{R}^d), \langle \cdot, \cdot \rangle_{L^2(\mathbb{R}^d)})$

Inner Product: $\left\langle \zeta, \hat{\zeta} \right\rangle_{g(\rho)} = \int_{\mathbb{R}^d} \zeta(x)\hat{\zeta}(x)dx, \quad \zeta, \hat{\zeta} \in T_\rho L^2(\mathbb{R}^d)$

If ρ is smooth, its tangent space is:

$$\{\zeta_1, \zeta_2, \cdots, \zeta_p\} = \{\partial_{\theta_1}\rho, \partial_{\theta_2}\rho, \cdots, \partial_{\theta_p}\rho\}$$

The L^2 - derivative of the f, w.r.t ρ is

$$\partial_\rho^g f = \partial_\rho f(\rho)$$

$$\Rightarrow \qquad \eta_{L^2}^{nat} = \underset{\eta \in \mathbb{R}^p}{\operatorname{argmin}} \left\| \partial_\rho f + \sum_{i=1}^{p} \eta_i \zeta_i \right\|_{L^2(\mathbb{R}^d)}^2 \qquad (\partial_\theta \rho)\eta = -\partial_\rho f$$

W_2 Natural Gradient Descent Formulation

Metric space: $(M, g) = (P_2(\mathbb{R}^d), W_2)$

$P_2(\mathbb{R}^d)$ is the set of Borel probability measures with finite second moments.

Wasserstein metric:

$$W_2(\rho_1, \rho_2) = \left(\inf_{\pi \in \Gamma(\rho_1, \rho_2)} \int_{\mathbb{R}^{2d}} |x - y|^2 d\pi(x, y) \right)^{\frac{1}{2}} \qquad \pi : \mathbb{R}^{2d} \to \mathbb{R}$$

Two Important Ingredients:

1. W_2 Tangent Vectors:

$$\left\{ v_1, v_2, \cdots, v_p \right\} = \left\{ \partial^W_{\theta_1} \rho, \partial^W_{\theta_2} \rho, \cdots, \partial^W_{\theta_p} \rho \right\}$$

$$v_i = argmin_v \left\{ \|v\|^2_{L^2_{\rho(\theta)}(\mathbb{R}^d; \mathbb{R}^d)} : \quad - \nabla \cdot (\rho(\theta) v) = \zeta_i \right\}$$

2. The W_2 - derivative of the f, w.r.t ρ is $\partial^W_\rho f = \nabla \partial_\rho f$

W_2 Natural Gradient Descent Formulation

$$\eta_{W_2}^{nat} = \underset{\eta \in \mathbb{R}^p}{\operatorname{argmin}} \left\| \partial_\rho^W f + \sum_{i=1}^{p} \eta_i v_i \right\|_{L^2_\rho(\mathbb{R}^d;\mathbb{R}^d)}^2$$

Change Variable:

$$\tilde{v}_i = \sqrt{\rho}\, v_i$$

$$\Rightarrow \quad \tilde{v}_i = \underset{\tilde{v}}{\operatorname{argmin}} \left\{ \|\tilde{v}\|^2_{L^2(\mathbb{R}^d;\mathbb{R}^d)} : \; -\nabla \cdot \left(\sqrt{\rho(\theta)}\, \tilde{v} \right) = \zeta_i \right\}$$

$\tilde{v}_i$ Is the solution to this constrained least norm problem:

$$\min_v ||v||_2^2, s.t. Bv = \zeta_i = \partial_{\theta_i} \rho \quad \Leftrightarrow \quad BY = \partial_\theta \rho$$

Denote: $Y = (\tilde{v}_1, \tilde{v}_2, ..., \tilde{v}_p) = B^\dagger \partial_\theta \rho$ $\qquad B \leftrightarrow -\nabla \cdot (\sqrt{\rho}, \cdot)$

$$\Rightarrow \quad \eta_{W_2}^{nat} = \underset{\eta \in \mathbb{R}^p}{\operatorname{argmin}} \left\| \sqrt{\rho}\, \partial_\rho^W f + \sum_{i=1}^{p} \eta_i \tilde{v}_i \right\|_{L^2(\mathbb{R}^d;\mathbb{R}^d)}^2 = -Y^\dagger \left(\sqrt{\rho}\, \partial_\rho^W f \right)$$

W_2 & L^2 Natural Gradient Descent Summary

Denote: $Z = \partial_\theta \rho$ $\quad Y = B^\dagger Z$

Method #1:

$$\eta^{nat}_{L^2} = G_{L^2}(\theta)^{-1}\eta^{std} \qquad G_{L^2}(\theta) = Z^T Z$$

$$\eta^{nat}_{W^2} = G_{W^2}(\theta)^{-1}\eta^{std} \qquad G_{W^2}(\theta) = Y^T Y$$

Method #2:

$$\eta^{nat}_{W_2} = \underset{\eta\in\mathbb{R}^p}{\operatorname{argmin}} \left\| \sqrt{\rho}\, \partial^W_\rho f + \sum_{i=1}^{p} \eta_i \tilde{v}_i \right\|^2_{L^2(\mathbb{R}^d;\mathbb{R}^d)} = -Y^\dagger \left(\sqrt{\rho}\, \partial^W_\rho f\right)$$

$$\eta^{nat}_{L^2} = \underset{\eta\in\mathbb{R}^p}{\operatorname{argmin}} \left\| \partial_\rho f + \sum_{i=1}^{p} \eta_i \zeta_i \right\|^2_2 = -Z^\dagger \partial_\rho f$$

Numerical Methods to Compute W_2 and L^2 Natural Gradient Descent Direction

Adjoint-State Method For ρ implicitly given by some function g:

Step 0: Implicit constraint: $g(\rho(\theta), \theta) = 0 \quad g : \mathbb{R}^k \times \mathbb{R}^p \to \mathbb{R}^k$

Step 1: Take the first derivative with respect to θ:

$$\partial_\rho g \partial_\theta \rho + \partial_\theta g = 0$$

$$\Rightarrow \partial_\rho g \ \partial_\theta \rho = \partial_\rho g \ Z = -\partial_\theta g \qquad (Z = \partial_\theta \rho)$$

Step 2: For any vector h, solve for its adjoint variable λ:

$$\lambda^\top \partial_\rho g = h^\top \quad \Longleftrightarrow \quad (\partial_\rho g)^\top \lambda = h$$

Step 3: Then we can get:

$$h^T \partial_\theta \rho = \ h^\top Z = \lambda^\top \partial_\rho g \ Z = -\lambda^\top \partial_\theta g$$

Numerical Methods to Compute W_2 and L^2 Natural Gradient Descent Direction

Hutchinson Estimator

$h \in \mathbb{R}^k$ is a Hutchinson vector if h is a random vector with i.i.d. random entries with mean 0 and variance 1.

$$Z = \mathbb{E}\left[hh^\top Z\right]$$

With m Hutchinson vectors: $h_1, h_2, ..., h_m$

The m'th order Hutchinson estimation of matrix Z is defined as:

$$H_m(Z) = \frac{1}{m}\sum_{k=1}^{m} h_k h_k^\top Z$$

Numerical Methods to Compute W_2 and L^2 Natural Gradient Descent Direction

Adjoint-State Method and Hutchinson Estimator Combined

$$H_m(Z) = \frac{1}{m}\sum_{k=1}^{m} h_k h_k^\top Z$$

Let $\lambda_1, \lambda_2, ..., \lambda_m$ be the corresponding adjoint variables of $h_1, h_2, ..., h_m$

$$h^\top Z = \lambda^\top \partial_\rho g \; Z = -\lambda^\top \partial_\theta g$$

$$\Longrightarrow \quad H_m(Z) = -\frac{1}{m}\sum_{k=1}^{m} h_k \lambda_k^\top \partial_\theta g$$

where $\{\lambda_k\}$ satisfies:

$$(\partial_\rho g)^T \lambda_k = h_k$$

Numerical Methods to Compute W_2 and L^2 Natural Gradient Descent Direction

Adjoint-State Method and Hutchinson Estimator Combined

It is similar to approximate Y using adjoint state method and Hutchinson Estimator

$$Y = B^{\dagger} Z = B^{\dagger} \mathbb{E}\left(hh^{\top} Z\right) = \mathbb{E}\left(B^{\dagger} hh^{\top} Z\right)$$

$$H_m(Y) = \frac{1}{m} \sum_{k=1}^{m} B^{\dagger} h_k h_k^{\top} Z = \frac{1}{m} \sum_{k=1}^{m} B^{\dagger} h_k \lambda_k^{\top} \partial_{\theta} g$$

Implementation of Adjoint-State Method and Hutchinson Estimator

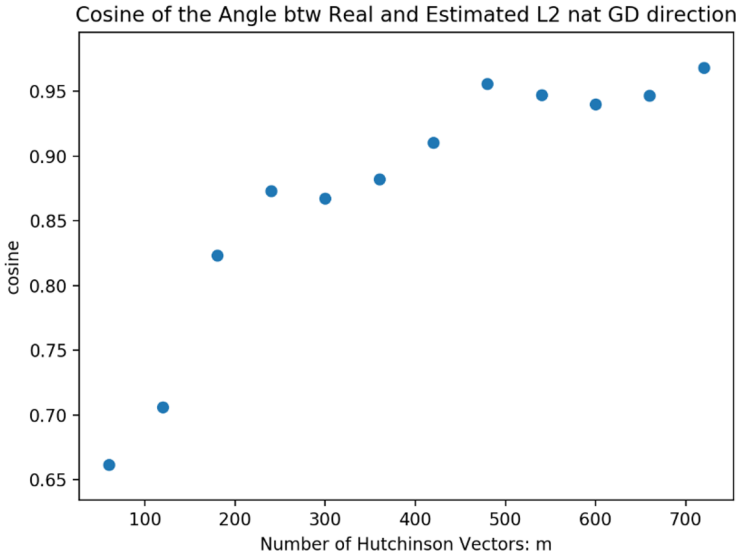
In Lorenz System

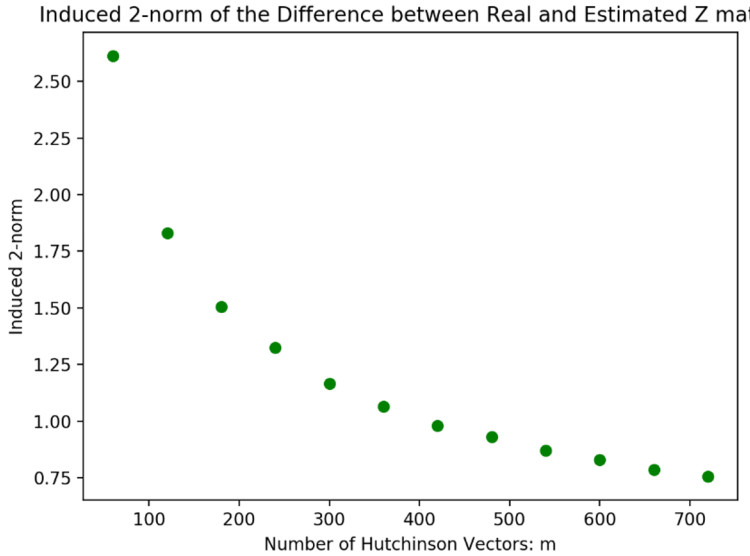
$$\frac{dx}{dt} = \sigma(y - x)$$

$$\frac{dy}{dt} = x(\rho - z) - y$$

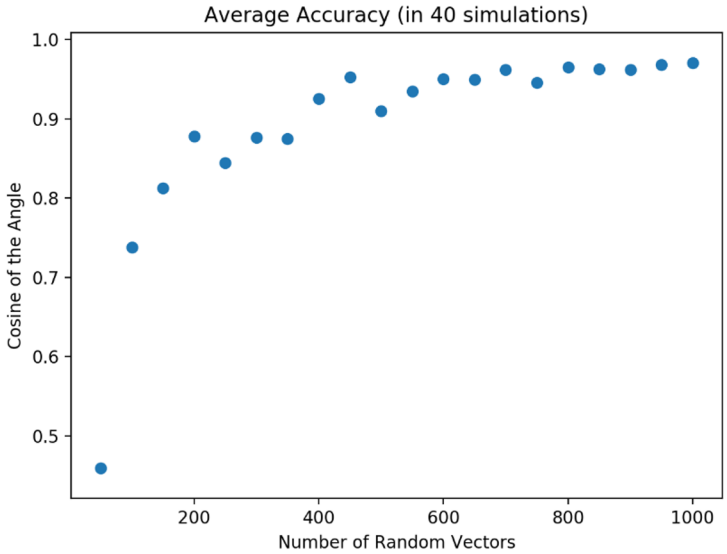
$$\frac{dz}{dt} = xy - \beta z$$

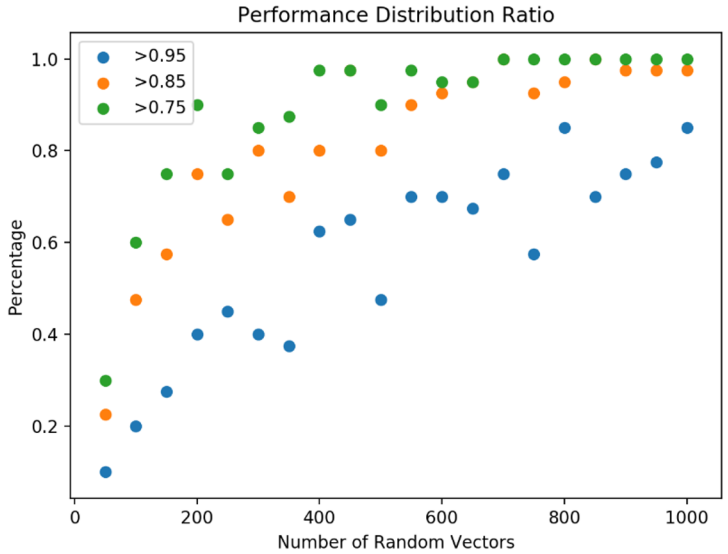
$$M(\theta)\rho - \rho = 0$$





Implementation of Adjoint-State Method and Hutchinson Estimator



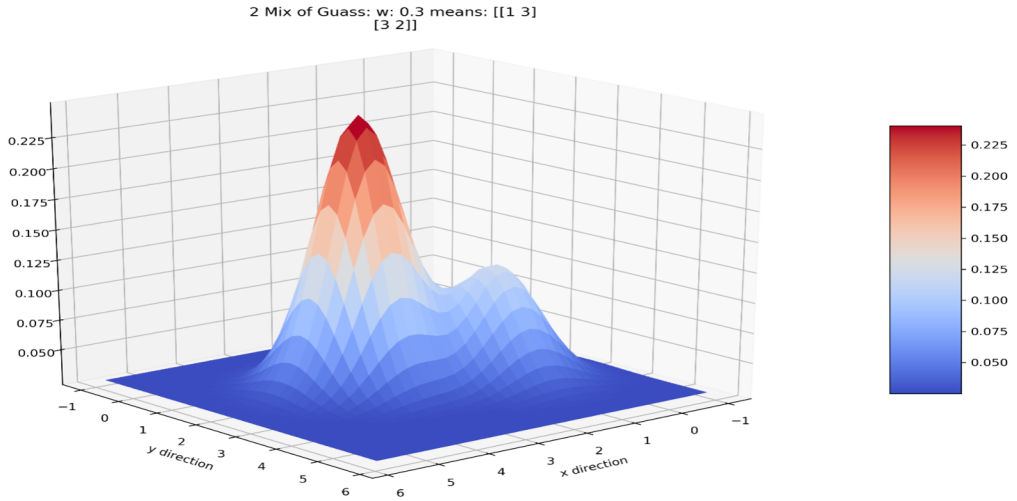


Implementation on two Mixtures of Gaussian Model in $\mathbb{R}^2$

Our Model: $$\rho(\theta, x) = wN(x; \mu_1, I) + (1 - w)N(x; \mu_2, I)$$

We fixed a truth model (reference model):

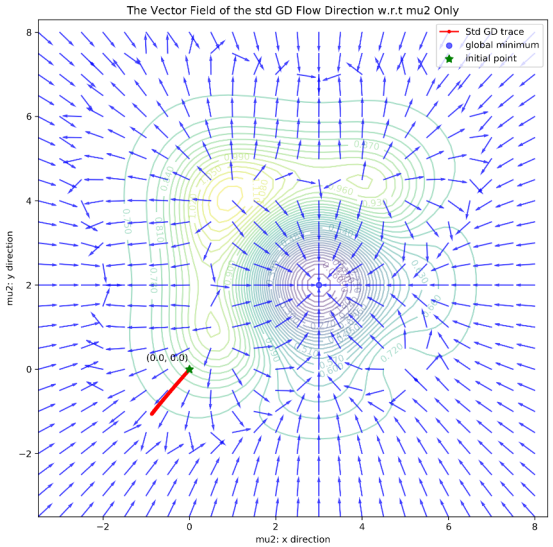
$$\rho(x) = 0.3N(x; [1, 3]^T, I) + 0.7N(x; [3, 2]^T, I)$$

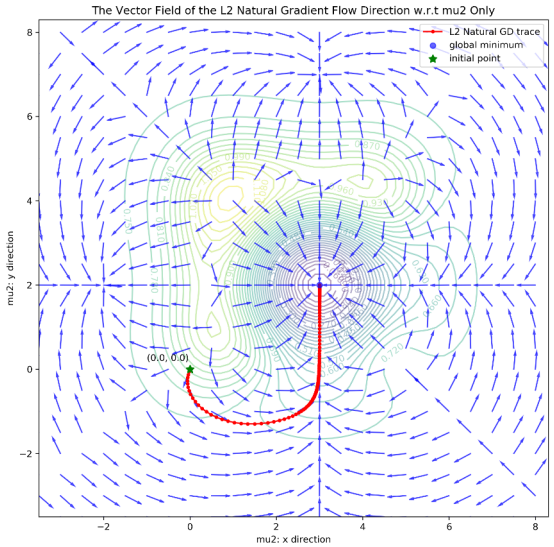


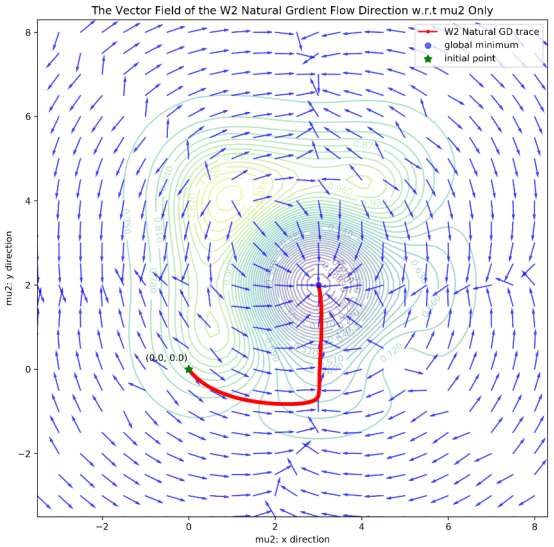
Implementation on two Mixtures of Gaussian Model in $\mathbb{R}^2$

Std GD Descent and Natural GD Descent Comparison

Fix w and mu2, only let mu1 be the varying parameters.



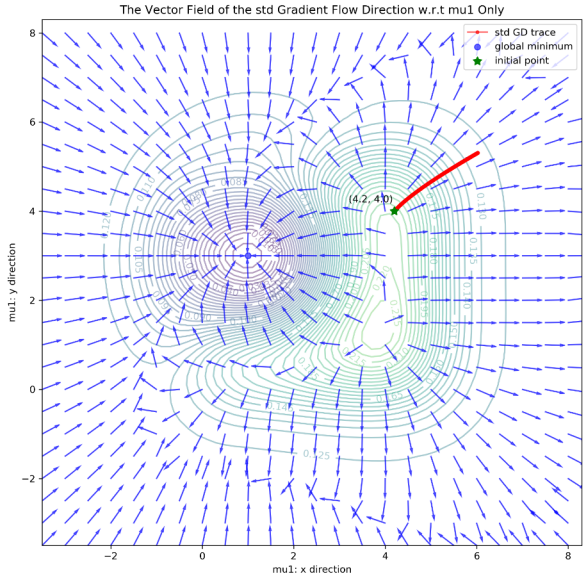


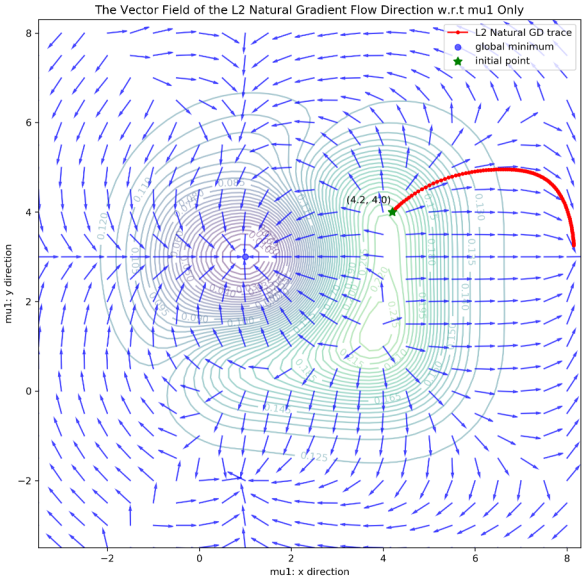


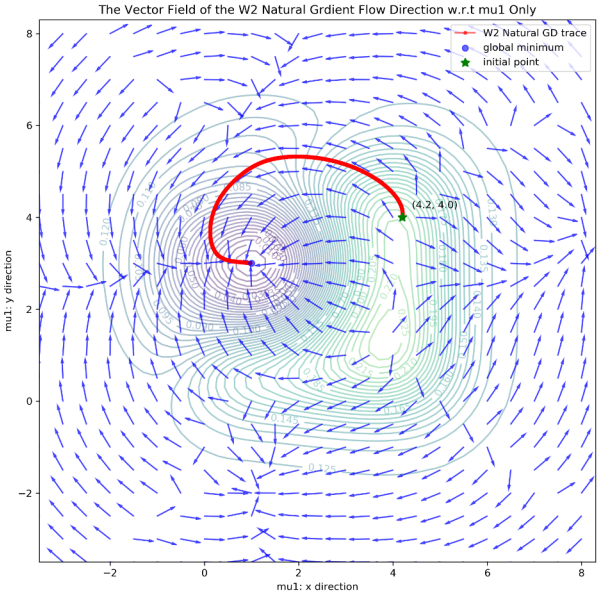
Implementation on two Mixtures of Gaussian Model in $\mathbb{R}^2$

Std GD Descent and Natural GD Descent Comparison

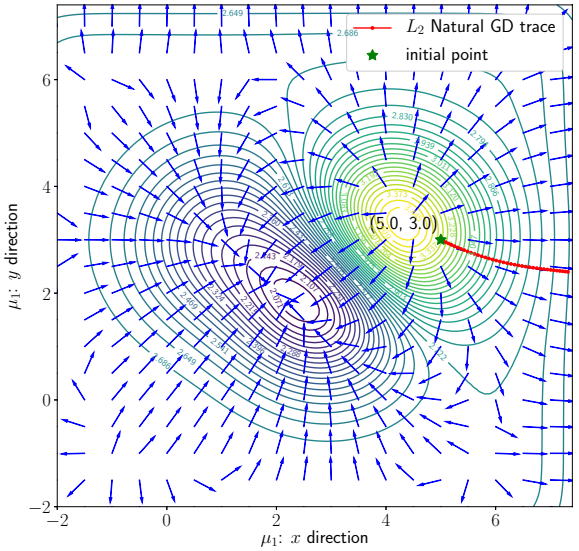
Natural gradient descent algorithm has a qualitative advantages. It is more likely to skip the local minimum and converge to global minimum in nonconvex situation.

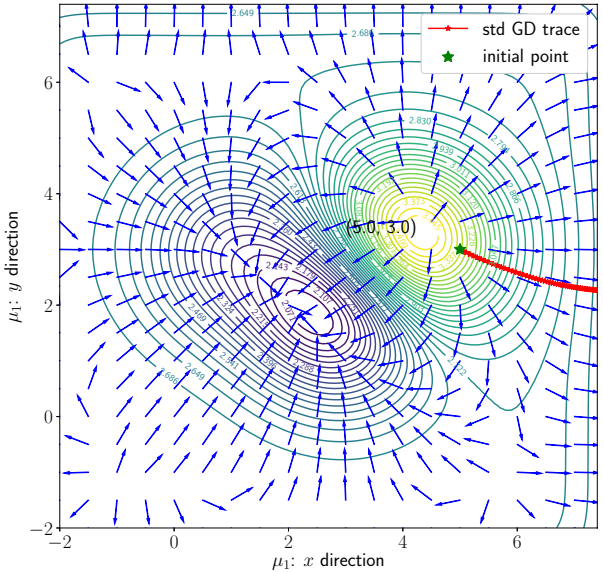


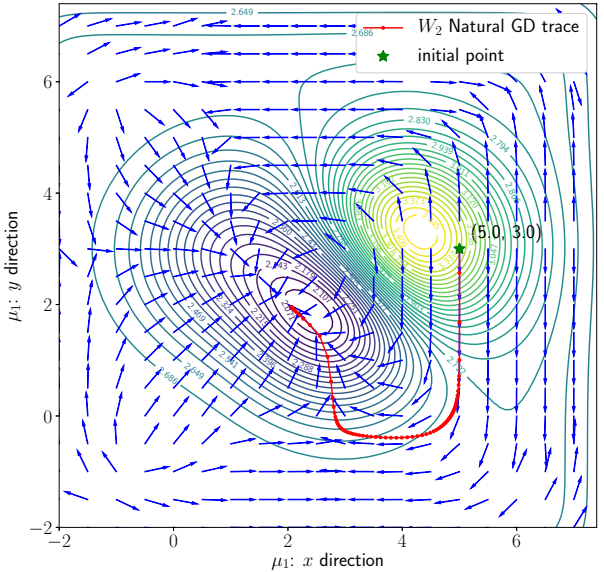




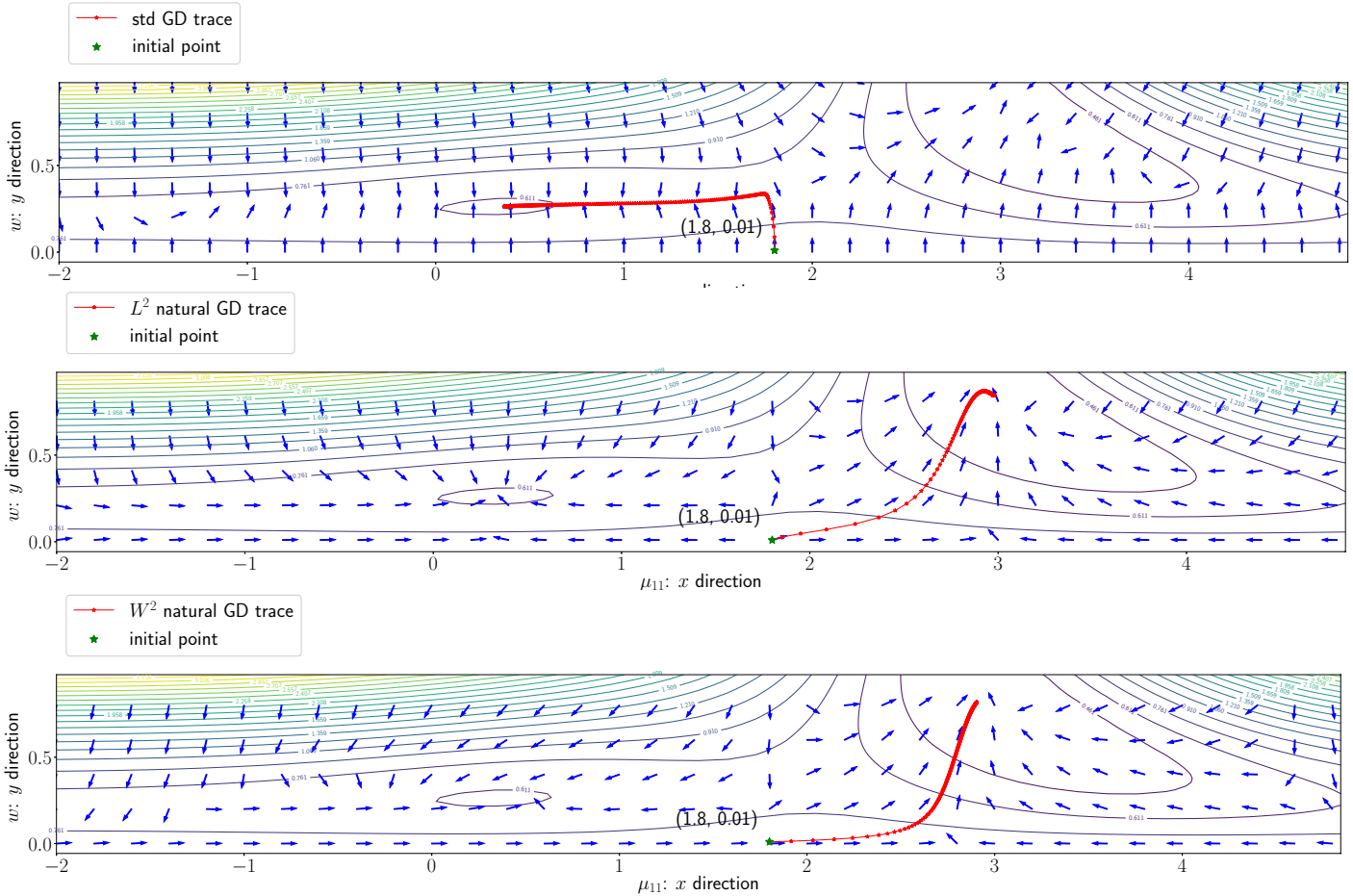
Implementation on two Mixtures of Gaussian Model in $\mathbb{R}^2$







Implementation on two Mixtures of Gaussian Model in $\mathbb{R}^2$



Fix mu2 and the second component mu_12 of the mean of the first mixture. Let the weight w and the first component of the first mean mu_11 be the varying parameters.

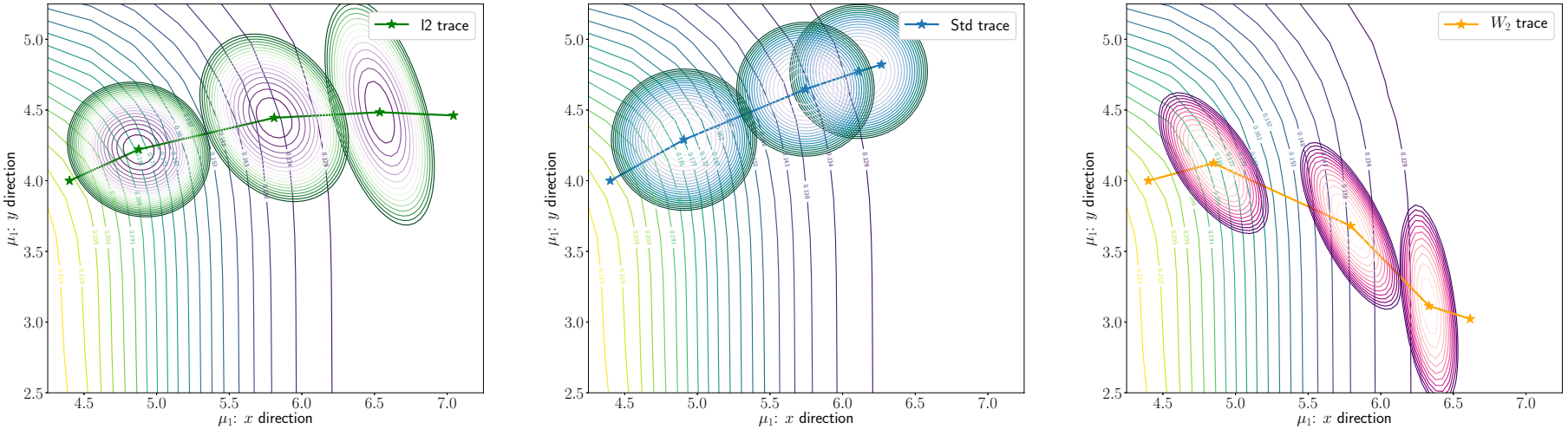
Natural gradient descent skips the local minimum and converges to the global minimum.

Implementation on two Mixtures of Gaussian Model in $\mathbb{R}^2$

$$
\begin{aligned}
\theta_{\text{std}}^{l+1} &= \operatorname*{argmin}_{\theta} \left\{ f(\rho(\theta)) + \frac{\|\theta - \theta^l\|^2}{2\tau} \right\} \\
&\approx \theta^l + \operatorname*{argmin}_{h} \left\{ f(\rho(\theta^l)) + \nabla_\theta f^\top h + \frac{1}{2\tau} h^\top h \right\} \\
\theta_{L^2}^{l+1} &= \operatorname*{argmin}_{\theta} \left\{ f(\rho(\theta)) + \frac{||\rho(\theta) - \rho(\theta^l)||_2^2}{2\tau} \right\} \\
&\approx \theta^l + \operatorname*{argmin}_{h} \left\{ f(\rho(\theta^l)) + \nabla_\theta f^\top h + \frac{1}{2\tau} h^\top \partial_\theta \rho^\top \partial_\theta \rho \; h \right\} \\
\theta_{W_2}^{l+1} &= \operatorname*{argmin}_{\theta} \left\{ f(\rho(\theta)) + \frac{W_2^2(\rho(\theta), \rho(\theta^l))}{2\tau} \right\} \\
&\approx \theta^l + \operatorname*{argmin}_{h} \left\{ f(\rho(\theta^l)) + \nabla_\theta f^\top h + \frac{1}{2\tau} h^\top Y^\top Y h \right\} \\
&= \theta^l + \operatorname*{argmin}_{h} \left\{ f(\rho(\theta^l)) + \nabla_\theta f^\top h + \frac{1}{2\tau} h^\top (\partial_\theta \rho)^\top (B^\dagger)^\top B^\dagger \partial_\theta \rho \; h \right\}
\end{aligned}
$$

The convergence of std, W2 and L2 natural gradient descent are different because in each step, they minimize different quadratic functions locally.

Implementation on two Mixtures of Gaussian Model in $\mathbb{R}^2$



(a) Standard gradient descent (b) L^2 natural gradient (c) W_2 natural gradient

Fig. 4: Reconstructing μ_1 using three different gradient descent algorithms for the two-mixture model starting from $[4.2, 4]$ (a)-(c) show the local quadratic models for first several iterations of the three different methods.