Minimax Optimization

Nonsmooth Composite Nonconvex-Concave

Linglingzhi Zhu

Department of Systems Engineering and Engineering Management

The Chinese University of Hong Kong

26 Nov 2022

Joint work with Jiajin Li and Anthony Man-Cho So.

The 1st ORSHK Young Researchers Workshop

Our Focus

Nonconvex-concave minimax problems of the form

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} F(x, y),$$

where

- $F : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ is nonconvex-concave
- $\mathcal{X} \subseteq \mathbb{R}^n$ is closed convex (possibly compact)
- $\mathcal{Y} \subseteq \mathbb{R}^d$ is convex compact

Gradient Descent Ascent (GDA)

- Gradient Descent Ascent (GDA):

$$x^{k+1} := x^k - \alpha_k \nabla_x F(x^k, y^k),$$
$$y^{k+1} := y^k + \tau_k \nabla_y F(x^{k+1}, y^k),$$

  where α_k and τ_k are the step sizes.

- Strongly-Concave: GDA can generate an ϵ-stationary solution with iteration complexity $\mathcal{O}(\epsilon^{-2})$ [Lin et al. 2020] — matching the optimal!

- Concave: GDA suffers from oscillation — diminishing step size strategies $\mathcal{O}(\epsilon^{-6})$ [Lin et al. 2020].

Gradient Descent Ascent (GDA)

- Gradient Descent Ascent (GDA):

$$x^{k+1} := x^k - \alpha_k \nabla_x F(x^k, y^k),$$
$$y^{k+1} := y^k + \tau_k \nabla_y F(x^{k+1}, y^k),$$

  where α_k and τ_k are the step sizes.

- Strongly-Concave:
  GDA can generate an ϵ-stationary solution with iteration complexity $\mathcal{O}(\epsilon^{-2})$ [Lin et al. 2020] — matching the optimal!

- Concave: GDA suffers from oscillation — diminishing step size strategies $\mathcal{O}(\epsilon^{-6})$ [Lin et al. 2020].

Gradient Descent Ascent (GDA)

- Gradient Descent Ascent (GDA):

$$
\begin{aligned}
x^{k+1} &:= x^k - \alpha_k \nabla_x F(x^k, y^k), \\
y^{k+1} &:= y^k + \tau_k \nabla_y F(x^{k+1}, y^k),
\end{aligned}
$$

  where α_k and τ_k are the step sizes.

- Strongly-Concave:
  GDA can generate an ϵ-stationary solution with iteration complexity $\mathcal{O}(\epsilon^{-2})$ [Lin et al. 2020] — matching the optimal!

- Concave: GDA suffers from oscillation — diminishing step size strategies $\mathcal{O}(\epsilon^{-6})$ [Lin et al. 2020].

Oscillation of GDA

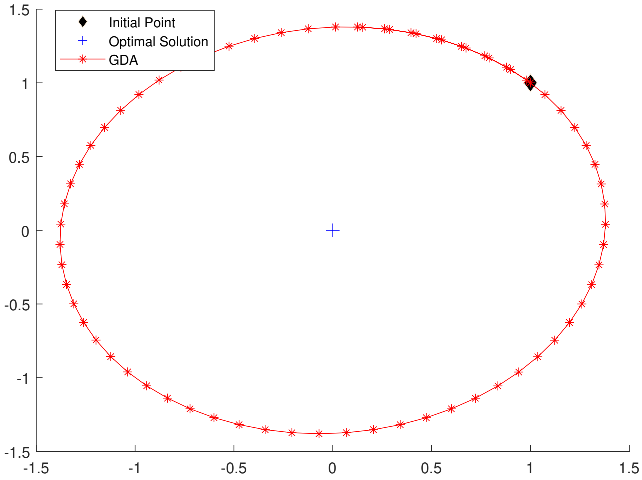


Figure: $F(x, y) = xy$, initial point: $(1, 1)$, optimal solution: $(0, 0)$

Smoothed GDA

- Iterative Scheme:

$$\begin{aligned} x^{k+1} &= x^k - \alpha_k[\nabla_x F(x^k, y^k) + \gamma(x^k - z^k)], \\ y^{k+1} &= \mathrm{proj}_{\mathcal{Y}}(y^k + \tau_k \nabla_y F(x^{k+1}, y^k)), \\ z^{k+1} &= z^k + \beta(x^{k+1} - z^k), \end{aligned}$$

where α_k and τ_k are the step sizes, β is the extrapolation parameter.

- Strongly Concave [Zhang et al. 2020]: Smoothed GDA can generate an ϵ-stationary solution with iteration complexity $\mathcal{O}(\epsilon^{-2})$ — matching the optimal!
- Concave: $\mathcal{O}(\epsilon^{-4})$ [Zhang et al. 2020] — best known result.

Smoothed GDA

- Iterative Scheme:

$$
\begin{aligned}
x^{k+1} &= x^k - \alpha_k[\nabla_x F(x^k, y^k) + \gamma(x^k - z^k)],\\
y^{k+1} &= \mathrm{proj}_{\mathcal{Y}}(y^k + \tau_k \nabla_y F(x^{k+1}, y^k)),\\
z^{k+1} &= z^k + \beta(x^{k+1} - z^k),
\end{aligned}
$$

  where α_k and τ_k are the step sizes, β is the extrapolation parameter.

- Strongly Concave [Zhang et al. 2020]:
  Smoothed GDA can generate an ϵ-stationary solution with iteration complexity $\mathcal{O}(\epsilon^{-2})$ — matching the optimal!

- Concave: $\mathcal{O}(\epsilon^{-4})$ [Zhang et al. 2020] — best known result.

Smoothed GDA

- Iterative Scheme:

$$\begin{aligned} x^{k+1} &= x^k - \alpha_k[\nabla_x F(x^k, y^k) + \gamma(x^k - z^k)], \\ y^{k+1} &= \mathrm{proj}_{\mathcal{Y}}(y^k + \tau_k \nabla_y F(x^{k+1}, y^k)), \\ z^{k+1} &= z^k + \beta(x^{k+1} - z^k), \end{aligned}$$

  where α_k and τ_k are the step sizes, β is the extrapolation parameter.

- Strongly Concave [Zhang et al. 2020]:
  Smoothed GDA can generate an ϵ-stationary solution with iteration complexity $\mathcal{O}(\epsilon^{-2})$ — matching the optimal!

- Concave: $\mathcal{O}(\epsilon^{-4})$ [Zhang et al. 2020] — best known result.

Motivation

- (Smoothed) GDA relies on **gradient Lipschitz** condition.
- (Rafique et al. 2021) has proposed an algorithm for general nonsmooth weakly convex-concave problems but suffers from the **slow iteration complexity** $\mathcal{O}(\epsilon^{-6})$.

Can we design a provably efficient algorithm to address nonsmooth nonconvex-concave (NNC-C) minimax problems, which matches the best known results for smooth case?

Problem Setup

- **(Primal Function)** $F(\cdot, y) := h_y \circ c_y$, where
  - $c_y : \mathbb{R}^n \to \mathbb{R}^m$ is C^1 and

$$\|\nabla c_y(x) - \nabla c_y(x')\| \leqslant L_c \|x - x'\| \quad \text{for all } x, x' \in \mathcal{X},$$

  - $h_y : \mathbb{R}^m \to \mathbb{R}$ is convex and Lipschitz.
- For example, $h_y = \|\cdot\|_p$, where $p = \{1, 2, +\infty\}$.

Problem Setup

- (**Dual Function**) $F(x,\cdot)$ is concave and C^1 on $\mathcal{Y}$ with $\nabla_y F(\cdot,\cdot)$ being L-Lipschitz continuous on $\mathcal{X} \times \mathcal{Y}$, i.e.,

$$\|\nabla_y F(x,y) - \nabla_y F(x',y')\| \leqslant L\|(x,y) - (x',y')\|$$

  for all $(x,y), (x',y') \in \mathcal{X} \times \mathcal{Y}$.

Applications

NNC-C minimax problem has attracted intense attention across **optimization** and **machine learning** communities.

- Adversarial Training
- Generative Adversarial Network (GAN)
- Distributionally Robust Optimization (DRO)

Applications

- Distributionally Robust Optimization (DRO):

$$\min_{x \in \mathcal{X}} \max_{Q \in \mathcal{U}(\mathbb{P}_N)} \mathbb{E}_{\xi \sim Q}[f(x;\xi)]$$

  - $\mathbb{P}_N$: empirical distribution;
  - $\mathcal{U}(\mathbb{P}_N)$: ambiguity set defined by a host of probability metrics, e.g., f-divergence, Wasserstein, etc

$$\mathcal{U}(\mathbb{P}_N) = \{Q : d(Q, \mathbb{P}_N) \leqslant r\}.$$

- Variation Regularized Wasserstein DRO:

$$\min_{\theta} g(\theta) := \mathbb{E}_{\mathbb{P}_N}[\ell(y, f_\theta(x))] + \rho \max_{i \in [N]} \|\nabla_x \ell(y_i, f_\theta(x_i))\|_p.$$

Smoothed PLDA

Smoothed **P**roximal **L**inear **D**escent **A**scent (Smoothed PLDA):

$$
\begin{aligned}
x^{k+1} &:= \operatorname*{argmin}_{x \in \mathcal{X}} \Big\{ h_{y^k}\left(c_{y^k}(x^k) + \nabla c_{y^k}(x^k)^\top (x - x^k)\right) \\
&\qquad\qquad + \frac{\lambda}{2}\|x - x^k\|^2 + \frac{r}{2}\|x - z^k\|^2 \Big\} \\
y^{k+1} &:= \operatorname{proj}_{\mathcal{Y}}\left(y^k + \alpha \nabla_y F(x^{k+1}, y^k)\right) \\
z^{k+1} &:= z^k + \beta(x^{k+1} - z^k)
\end{aligned}
$$

No available gradient information due to composite structure $h_y \circ c_y$.

Here, we invoke the proximal linear scheme for the primal update.

Main Results

Table 1: Comparison of the iteration complexities of smoothed PLDA proposed in this paper and other related methods under different settings for solving $\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} F(x, y)$.

	Primal Func.	**Dual Func.**	**Iter. Compl.**[1]	**Add. Asm.**
GDA	L-smooth	concave	$\mathcal{O}(\epsilon^{-6})$	$\mathcal{X} = \mathbb{R}^n$
Smoothed GDA	L-smooth	concave	$\mathcal{O}(\epsilon^{-4})$	—
PG-SMD	weakly-convex	concave	$\mathcal{O}(\epsilon^{-6})$	$\mathcal{X}$ bounded
This paper	nonsmooth composite	concave	$\mathcal{O}(\epsilon^{-4})$	—
GDA	L-smooth	strongly-concave	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{X} = \mathbb{R}^n$
Smoothed GDA	L-smooth	PŁ condition	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{Y} = \mathbb{R}^d$
This paper	nonsmooth composite	KŁ exponent $\theta = \frac{1}{2}$	$\mathcal{O}(\epsilon^{-2})$	—

Lipschitz-type Primal Error Bound Condition

Main Technical Results I

For any $k \geqslant 0$, it holds that

$$\|x^{k+1} - x_r(y^k, z^k)\| \leqslant \zeta \|x^k - x^{k+1}\|,$$

where $\zeta := \frac{2(r-L)^{-1}+(\lambda+L)^{-1}}{(\lambda+L)^{-1}} \left(\sqrt{\frac{2L}{\lambda+L}} + 1\right)$ and $x_r(y,z) := \operatorname*{argmin}_{x \in \mathcal{X}} F_r(x,y,z) := F(x,y) + \frac{r}{2}\|x - z\|^2$.

- Smooth case: Luo-Tseng error bound condition

$$\|x^{k+1} - x_r(y^k, z^k)\| \leqslant \zeta \|x^k - \underbrace{\mathrm{proj}_{\mathcal{X}}(x^k - c\nabla_x F_r(x^k, y^k, z^k))}_{x^{k+1}}\|,$$

Lyapunov Function

Define a Lyapunov function as

$$\Phi_r(x,y,z) := \underbrace{F_r(x,y,z) - d_r(y,z)}_{\text{Primal Descent}} + \underbrace{p_r(z) - d_r(y,z)}_{\text{Dual Ascent}} + \underbrace{p_r(z)}_{\text{Proximal Descent}}.$$

- $d_r(y,z) := \min_{x \in \mathcal{X}} F_r(x,y,z)$;
- $p_r(z) := \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} F_r(x,y,z)$;

Sufficient Decrease Property

Proposition

$r \geqslant 3L$, $\lambda \geqslant L$, $\beta \leqslant \min\left\{\frac{1}{28}, \frac{(r-L)^2}{32\alpha r(r+L)^2}\right\}$ *and* $\alpha \leqslant \min\left\{\frac{1}{10L}, \frac{1}{4L\zeta^2}\right\}$. *Then for any* $k \geqslant 0$,

$$\Phi_r^k - \Phi_r^{k+1} \geqslant \frac{\lambda}{16}\|x^k - x^{k+1}\|^2 + \frac{1}{8\alpha}\|y^k - y_+^k(z^k)\|^2 + \frac{4r}{7\beta}\|z^k - z^{k+1}\|^2 - \boxed{28r\beta\|x_r^*(z^k) - x_r(y_+^k(z^k), z^k)\|^2},$$

where $y_+(z) := \mathrm{proj}_{\mathcal{Y}}\left(y + \alpha\nabla_y F_r(x_r(y,z), y, z)\right)$ *and*

$$x_r^*(z) := \operatorname*{argmin}_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} F_r(x, y, z).$$

KŁ Exponent θ for the Dual Function

Kurdyka-Łojasiewicz (KŁ) Exponent

For any fixed $x \in \mathcal{X}$, there exist $\mu > 0$ and $\theta \in [0, 1)$ such that

$$\operatorname{dist}(0, -\nabla_y F(x, y) + \partial \iota_{\mathcal{Y}}(y)) \geqslant \mu \left(\max_{y' \in \mathcal{Y}} F(x, y') - F(x, y) \right)^{\theta},$$

for any $y \in \mathcal{Y}$.

Generalization of the **strong concavity** of the dual function.

Dual Error Bound Condition

Main Technical Results II

Suppose that the dual function satisfies KŁ property with exponent $\theta \in [0, 1)$. Then

$$\|x_r^*(z) - x_r(y_+(z), z)\| \leqslant \omega \|y - y_+(z)\|^{\frac{1}{2\theta}},$$

where $\omega := \frac{\sqrt{2}}{\sqrt{r-L}} \left(\frac{(1+\alpha L)(r-L+2\alpha L(r+L))}{\alpha\mu(r-L)} \right)^{\frac{1}{2\theta}}$.

Explicitly control the trade-off between the decrease in the primal and the increase in the dual.

Stationarity Concept

Definition

The pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is an ϵ-game stationary point (ϵ-GS) if [1]

$$\|\nabla_x d_r(y, x)\| \leqslant \epsilon \quad \text{and} \quad \operatorname{dist}(0, -\nabla_y F(x, y) + \partial \iota_{\mathcal{Y}}(y)) \leqslant \epsilon.$$

With the aid of our newly developed dual error bound condition, we can clarify the relationship among various stationarity concepts quantitatively.

[1] $\|\nabla_x d_r(y, x)\|$ reduces to $\operatorname{dist}(0, -\nabla_x F(x, y) + \partial \iota_{\mathcal{X}}(x))$ for the smooth case.

Stationarity Concept

Definition

The pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is an ϵ-game stationary point (ϵ-GS) if [1]

$$\|\nabla_x d_r(y, x)\| \leqslant \epsilon \quad \text{and} \quad \operatorname{dist}(0, -\nabla_y F(x, y) + \partial \iota_{\mathcal{Y}}(y)) \leqslant \epsilon.$$

With the aid of our newly developed dual error bound condition, we can clarify the relationship among various stationarity concepts quantitatively.

[1] $\|\nabla_x d_r(y, x)\|$ reduces to $\operatorname{dist}(0, -\nabla_x F(x, y) + \partial \iota_{\mathcal{X}}(x))$ for the smooth case.

Quantitative Results for Stationarities

- The point $x \in \mathcal{X}$ is an ϵ-**optimization stationary**[2] (ϵ-OS) if

$$\| \operatorname{prox}_{\frac{1}{r} f + \iota_{\mathcal{X}}}(x) - x \| \leqslant \epsilon.$$

Main Technical Results III

Suppose that the pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is ϵ-GS. Then, x is $\mathcal{O}(\epsilon^{\frac{1}{2}})$-OS. Moreover, if the dual function satisfies KŁ property with exponent θ, then x is $\mathcal{O}(\epsilon^{\min\{1, \frac{1}{2\theta}\}})$-OS.

[2] It reduce to $\mathrm{d}(0, \partial(f + \iota_{\mathcal{X}})(x))$ for the smooth case with $f := \max_{y \in \mathcal{Y}} F(\cdot, y)$.

Main Theorem — Iteration Complexity

Suppose that $r \geqslant 3L$, $\lambda \geqslant L$, $\beta \leqslant \min\left\{\frac{1}{28}, \frac{(r-L)^2}{32\alpha r(r+L)^2}\right\}$ and $\alpha \leqslant \min\left\{\frac{1}{10L}, \frac{1}{4L\zeta^2}\right\}$. Then for any $k \geqslant 0$,

- General concave: there exists a $k \in [K]$ such that $\left(x^{k+1}, y^{k+1}\right)$ is an $\mathcal{O}(K^{-\frac{1}{4}})$-game stationary if $\beta \leqslant K^{-\frac{1}{2}}$.
- KŁ exponent $\theta \in (\frac{1}{2}, 1)$: there exists a $k \in [K]$ such that $\left(x^{k+1}, y^{k+1}\right)$ is an $\mathcal{O}(K^{-\frac{1}{4\theta}})$-game stationary if $\beta \leqslant K^{-\frac{2\theta-1}{2\theta}}$.
- KŁ exponent $\theta \in [0, \frac{1}{2}]$: there exists a $k \in [K]$ such that $\left(x^{k+1}, y^{k+1}\right)$ is an $\mathcal{O}(K^{-\frac{1}{2}})$-game stationary if $\beta = \mathcal{O}(1)$.

Numerical Results

Recall the variation regularized Wasserstein DRO:

$$\min_{\theta} g(\theta) := \mathbb{E}_{\mathbb{P}_N}[\ell(y, f_\theta(x))] + \rho \max_{i\in[N]} \|\nabla_x \ell(y_i, f_\theta(x_i))\|_p. \qquad (1)$$

- $\ell : \mathbb{R} \to \mathbb{R}$ is the loss function;
- $f_\theta : \mathbb{R}^d \to \mathbb{R}$ is the feature mapping;
- $\{(x_i, y_i)\}_{i=1}^N$ is the training dataset and $p = \{1, 2, +\infty\}$;
- closed connection with the Lipschitz constant of deep neural networks;

Numerical Results

Recall the variation regularized Wasserstein DRO:

$$\min_{\theta} g(\theta) := \mathbb{E}_{\mathbb{P}_N}[\ell(y, f_\theta(x))] + \rho \max_{i\in[N]} \|\nabla_x \ell(y_i, f_\theta(x_i))\|_p. \tag{1}$$

- $\ell : \mathbb{R} \to \mathbb{R}$ is the loss function;
- $f_\theta : \mathbb{R}^d \to \mathbb{R}$ is the feature mapping;
- $\{(x_i, y_i)\}_{i=1}^N$ is the training dataset and $p = \{1, 2, +\infty\}$;
- closed connection with the Lipschitz constant of deep neural networks;

Key Difficulties

- It is challenging for calculating the subdifferential set of the pointwise supremum of an arbitrary family (possibly not differentiable) of (weakly) convex functions.
- Minimax reformulation technique:

$$\min_{\theta} \max_{w \in \Delta_N} \mathbb{E}_{\mathbb{P}_N}[\ell(y, f_\theta(x))] + \rho \sum_{i=1}^{N} w_i \|\nabla_x \ell(y_i, f_\theta(x_i))\|_p, \quad (2)$$

which can be recast into the nonsmooth nonconvex-concave minimax problem.

Key Difficulties

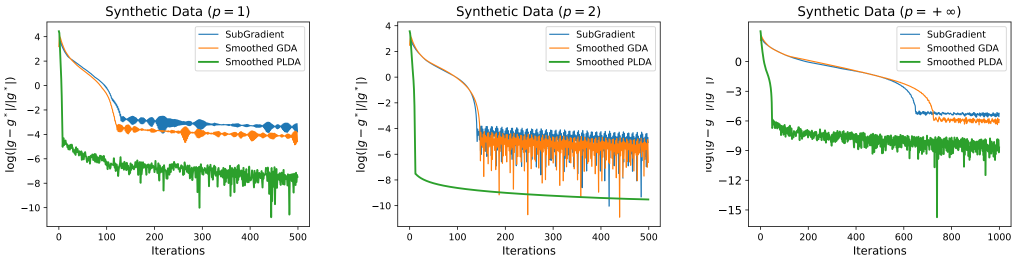
- It is challenging for calculating the subdifferential set of the pointwise supremum of an arbitrary family (possibly not differentiable) of (weakly) convex functions.
- Minimax reformulation technique:

$$\min_{\theta} \max_{w \in \Delta_N} \mathbb{E}_{\mathbb{P}_N} [\ell(y, f_\theta(x))] + \rho \sum_{i=1}^{N} w_i \|\nabla_x \ell(y_i, f_\theta(x_i))\|_p, \quad (2)$$

which can be recast into the nonsmooth nonconvex-concave minimax problem.

Linear Regression

Consider a simple case — the quadratic loss function with linear feature mapping, i.e., $\ell(y, f_\theta(x)) = \frac{1}{2}(y - \theta^\top x)^2$

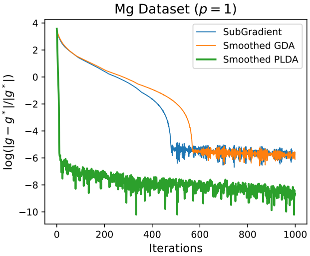


(a) $p = 1$ (b) $p = 2$ (c) $p = +\infty$

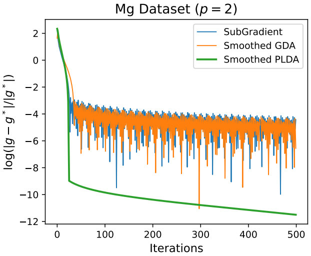
Figure: Compare the convergence behaviours of smoothed PLDA with subgradient and smoothed GDA on both synthetic and real world datasets.

Linear Regression

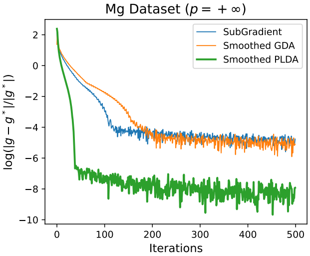
Consider a simple case — the quadratic loss function with linear feature mapping, i.e., $\ell(y, f_\theta(x)) = \frac{1}{2}(y - \theta^\top x)^2$



(a) $p = 1$



(b) $p = 2$

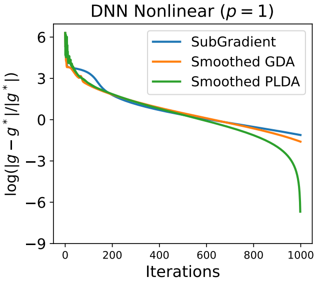


(c) $p = +\infty$

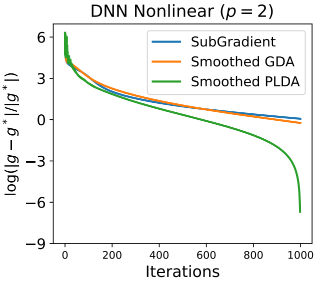
Figure: Compare the convergence behaviours of smoothed PLDA with subgradient and smoothed GDA on both synthetic and real world datasets.

Deep Neural Network

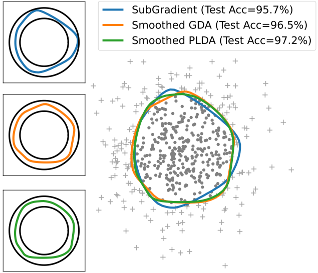
Here, $\ell(\cdot, \cdot)$ is the cross-entropy loss and $f_\theta(\cdot)$ is the feature mapping generated by a neural network with 2 hidden layers of size 5 and use the exponential linear unit (ELU) as the activation function.



(a) $p = 1$



(b) $p = 2$



(c) Decision boundary

Thank you for listening!

Linglingzhi Zhu

llzzhu@se.cuhk.edu.hk