Concave Conjugacy of Self-paced Learning

Shiqi Liu Instructor: Deyu Meng Cooperative partner: Zilu Ma Haocheng Wang

Research Training Presentation , 2016

Outline

1. Motivation
   - The Basic Problem That We Studied
2. Model Analysis
   - Equivalence
   - Approach
   - Previous Work
3. Construction
4. Self-paced Curriculum Learning

Outline

1. Motivation
   - The Basic Problem That We Studied
2. Model Analysis
   - Equivalence
   - Approach
   - Previous Work
3. Construction
4. Self-paced Curriculum Learning

SPL model

symbol description

- Data set $D = (x^i, y^i)_{i=1}^n$
- Decision function $p(x, w)$, w model parameter
- Loss function $L^i(y^i, p(x^i, w))$, $l = (L^1, \cdots, L^n)^T$
- λ a parameter controlling the learning space

The SPL model contains a weighted loss term $\langle v, l\rangle$, a model regularizer $\phi(w)$ and a self-paced regularizer $f(v, \lambda)$ imposed on sample weights, expressed as

$$\inf_{w, v\in[0,1]^n} E(w, v; \lambda) = \inf_{w, v\in[0,1]^n} \{\langle v, l\rangle + f(v, \lambda) + \phi(w)\}$$

Concave Conjugate vs. Convex Conjugate

The **concave conjugate** of function $g(v)$ is defined by the following

$$g^*(l) = \inf_{v \in R^n} \{\langle v, l \rangle - g(v)\}$$

The **convex conjugate** of $f(v)$[1] is defined by

$$f^*(l) = \sup_{v \in R^n} \{\langle v, l \rangle - f(v)\}$$

According to Fenchel's paper, $g^*(l)$ has the following properties

- upper semi-continuous
- concave

[1] Let $f(v) = -g(v)$; the concave conjugate and convex conjugate has the following relation

$$g^*(l) = -f^*(-l)$$

Outline

1 Motivation
- The Basic Problem That We Studied

2 Model Analysis
- Equivalence
- Approach
- Previous Work

3 Construction

4 Self-paced Curriculum Learning

Equivalence of SPL model

Latent Object Function and Its Properties

For concise presenting, we disregard λ

$$\inf_{w,v\in[0,1]^n} E(w,v) = \inf_{w,v\in[0,1]^n} \{\langle v,l\rangle + f(v) + \phi(w)\}$$

$$= \inf_{w}\{\phi(w) + \inf_{v\in[0,1]^n}\{\langle v,l\rangle + f(v)\}\} = \inf_{w}\{\phi(w) + g^*(l(w))\}$$

We call $F(l) = g^*(l)$ the latent object function, and SPL model turns into

$$\inf_{w,v\in[0,1]^n} E(w,v) = \inf_{w}\{\phi(w) + F(l(w))\}$$

$F(l)$ has the following properties

- upper semi-continuous
- concave
- increasing

Equivalence Class of SPL regularizer

Assumption 2 $f(v)$ satisfies the following properties on $[0,1]^n$

1. $f(v)$ is convex
2. $f(v)$ is lower semi-continuous
3. $int(dom\, f(v)) \cap int([0,1]^n) \neq \emptyset$

Outline

1 Motivation
- The Basic Problem That We Studied

2 Model Analysis
- Equivalence
- Approach
- Previous Work

3 Construction

4 Self-paced Curriculum Learning

Solving the SPL model

Alternative Optimization Strategy

Optimize $\inf\limits_{w,v\in[0,1]^n} v, w$ respectively, and suppose $f(v)$ satisfies **Assumption 2.**

in

$$E(w,v) = \inf_{w,v\in[0,1]^n}\{\langle v,l\rangle + f(v) + \phi(w)\}$$

$$= \inf_{w}\{\phi(w) + \inf_{v\in[0,1]^n}\{\langle v,l\rangle + f(v)\}\} = \inf_{w}\{\phi(w) + g^*(l(w))\}$$

According to the theorem in Convex Analysis, the following regime of AOS step can be derived.

- update v

$$v^i = \arg\inf_{v\in[0,1]^n} E(w^{i-1},v) = \arg\inf_{v\in[0,1]^n}\{\langle v,l\rangle + f(v)\} = \partial F(l(w^{i-1}))$$

- update w

$$w^i = \arg\inf_{w} E(w,v^i) = \arg\inf_{w}\{< v^i, l(w) > +\phi(w)\}$$

Outline

1 Motivation
- The Basic Problem That We Studied

2 Model Analysis
- Equivalence
- Approach
- Previous Work

3 Construction

4 Self-paced Curriculum Learning

Solving the SPL model

Majorization Minimization vs. Alternative Optimization Strategy

- In Deyu Meng's paper, the equivalence of Majorization Minimization and Alternative Optimization Strategy implemented on SPL model was proven by constructing the surrogate function[2]

$$Q_\lambda(w|w^*) = F_\lambda(l(w^*)) + \nabla F_\lambda(l(w^*))(l(w) - l(w^*))$$

[2] in case that $F_\lambda(l)$ is differentiable

Latent Object function

An approach to design SP-regularizer: First, design $v(l)$ satisfying $v_i(l)$ decrease with respect to l_i and

$$\lim_{l_i \to 0} v_i(l) = 1 \quad \lim_{l_i \to +\infty} v_i(l) = 0$$

and one could put the loss prior of different joint loss in it (Recommend that $v_i(l) = 1 \forall l_i < 0$);

Then, integrate $v(l)$ to obtain $F(l) = g^*(l)$; Last, $f(v, \lambda) = -\lambda g^{**}(v) = -\lambda g(v)$ and if $F(l)$ is strictly concave there is a shortcut that one can obtain $l(v)$ by calculating the inverse function of $v(l)$ then $g(v) = \langle v, l(v) \langle -F(l(v))$

SP-regularizer

Another approach to design SP-regularizer also can be derived by disregarding $\lim_{l_i \to 0} v(\lambda, l) = 1$.

First, design $f(v)$; let $f(v)$ satisfy

1. $int(dom\, f(v)) \cap (0,1)^n \neq \emptyset$ and $0, \mathbf{1} \in cl(dom\, f(v))$
2. $f(v)$ is convex, differentiable, lower semi-continuous in $v \in [0,1]^n$;

Then, obtain $l(v) = \partial(-f(v)) = \partial g(v)$ and calculate its inverse function $v(l)$; Last, integrate $v(l)$ to obtain $F(l) = g^*(l)$ or calculate $F(l) = g^*(l) = \langle v(l), l \langle + f(v(l))$;

$f(v, \lambda) = \lambda f(v)$, $F_\lambda(l) = \lambda F(\lambda^{-1} l)$ and $v(\lambda, l) = v(\lambda^{-1} l)$

Age parameter

The simplest SP-regularizer is $\lambda f(v)$. One can find that most of SP-regularizers, commonly appearing in SPL, can be generated in this way

The reason why it works is the following.

Let $g(v) = -f(v)$ and let the concave conjugate of $g^*(l) = F(l)$

$$F_\lambda(l) = (\lambda g(v))^* = \inf_{v \in [0,1]^n} \{\langle v, l \rangle - \lambda g(v)\}$$

$$= \lambda \inf_{v \in [0,1]^n} \{< v, \lambda^{-1} l > -g(v)\} = \lambda F(\lambda^{-1} l)$$

Since $g(v)$ is strictly concave, $F(l)$ is differentiable and the original $v^*(l) = \nabla F(l)$. It yields,

$$v^*(\lambda, l) = \nabla_l F_\lambda(l) = \lambda \nabla_l F(\lambda^{-1} l) = v^*(\lambda^{-1} l)$$

Age parameter

Thus, $v_i^*(\lambda, l)$ increase with respect to λ, and it holds that $\forall i \in \{1, 2, \cdots, n\}$, $\lim_{\lambda \to 0} v_i^*(\lambda, l) = \lim_{\lambda \to 0} v_i^*(\lambda^{-1} l) = 0$ and $\lim_{\lambda \to +\infty} v_i^*(\lambda, l) = \lim_{\lambda \to +\infty} v_i^*(\lambda^{-1} l) = 1$

One can also consider the hypograph of $F_\lambda(l)$, the set of points lying on or below its graph

$$hyp\ F_\lambda(l) = \{(l, u) : l \in R^n, u \in R, u \leq \lambda F(\lambda^{-1} l)\} = \lambda\ hyp\ F(l)$$

As a result, the geometric interpretation of the effect of age parameter here is that it enlarges the hypograph of $F(l)$ by multiplying λ .

Curriculum Region

In [Self-paced Curriculum Learning, LuJiang, Deyu Meng] paper, they put the prior knowledge into model by the constraints on the feasible region of v denoted by Ψ.

Suppose $f(v)$ satisfying **Assumption 2**, let $F(l) = \inf_{v\in[0,1]^n}\{\langle v, l\rangle + f(v)\}$ denote the concave conjugate of $-f(v)$.

In this case, we have

$$F^{new}(l) = \inf_{v\in[0,1]^n\cap\Psi}\{\langle v, l\rangle + f(v)\} = \inf_{v\in[0,1]^n}\{\langle v, l\rangle + f(v) - \delta(v|\Psi)^3\}$$

$$\inf_{w,v\in[0,1]^n\cap\Psi} E(w, v) = \inf_{w}\{\phi(w) + F^{new}(l(w))\}$$

[3] $\delta(v|\Psi) = 0\ \forall v \in \Psi \quad \delta(v|\Psi) = -\infty\ \forall v \notin \Psi$

Sup Convolution

sup convolution

$$(f \oplus g)(v) = \sup_{v^1+v^2=v} \{f(v^1) + g(v^2)\}$$

Theorem 0 Let $g_1, \cdots, g_m$ be proper concave function on R^n. Then

$$(g_1 \oplus \cdots \oplus g_m)^* = g_1^* + \cdots + g_n^*$$

$$(clg_1 + \cdots + clg_m)^* = cl(g_1^* \oplus \cdots \oplus g_m^*)$$

If sets, the relative interior of $(dom\ g_i), i = 1, \cdots, m$, have a point in common, the closure operation can be omitted from the second formula, and

$$(g_1 + \cdots + g_m)^* = g_1^* \oplus \cdots \oplus g_m^*$$

New Latent Object Function

$$F^{new}(l) = \inf_{v \in [0,1]^n \cap \Psi} \{\langle v, l \rangle + f(v)\} = \inf_{v \in [0,1]^n} \{\langle v, l \rangle + f(v) - \delta(v|\Psi)^4\}$$
$$= (-f(v) + \delta(v|\Psi))^* = (F \oplus \delta^*(\cdot|\Psi))(l)$$

Theorem 4 Suppose $f(v)$ is essential strictly convex and satisfies **Assumption 2**. Suppose we have knowledge of the weight variable v, denoted by $v^T k \geq 0$ corresponding to $\Psi = \{v | v^T k \geq 0\}$. If $\Psi \cap dom\, f \cap [0,1]^n \neq dom\, f \cap [0,1]^n$ $int(\Psi) \cap int(dom\, f) \cap int([0,1]^n) \neq \emptyset$

Then

$$\nabla F^{new}(l)^T k = v^{new}(l)^T k \geq 0$$

[4] $\delta(v|\Psi) = 0\ \forall v \in \Psi \quad \delta(v|\Psi) = -\infty\ \forall v \notin \Psi$

Suppose $f(v)$ is essential strictly convex and satisfies **Assumption 2**. Curriculum region $\Psi = \{v|v^T k \geq 0\}$ satisfies $\Psi \cap dom\, f \cap [0,1]^n \neq dom\, f \cap [0,1]^n$ $int(\Psi) \cap int(dom\, f) \cap int([0,1]^n) \neq \emptyset$

$$F^{new}(l) = \inf_{v \in [0,1]^n \cap \Psi} \{\langle v, l\rangle + f(v)\} = \inf_{v \in [0,1]^n} \{\langle v, l\rangle + f(v) - \delta(v|\Psi)^5\}$$

$$= (-f(v) + \delta(v|\Psi))^* = (F \oplus \delta^*(\cdot|\Psi))(l) = (F \oplus \delta(\cdot|\Psi^{\circ 6}))(l)$$

$$= \sup_{l^1 + l^2 = l} \{F(l^1)) + \delta(l^2|\Psi^\circ)\} = \sup_{l^1 \in l - \Psi^\circ} F(l^1) = \sup_{l^1 \in l - ray_k} F(l^1)$$

[5] $\delta(v|\Psi) = 0\ \forall v \in \Psi \quad \delta(v|\Psi) = -\infty\ \forall v \notin \Psi$
[6] $\Psi^\circ = \{l | \forall v \in \Psi\ \langle v, l\rangle \geq 0\} = \{\beta k | \beta \geq 0\}$

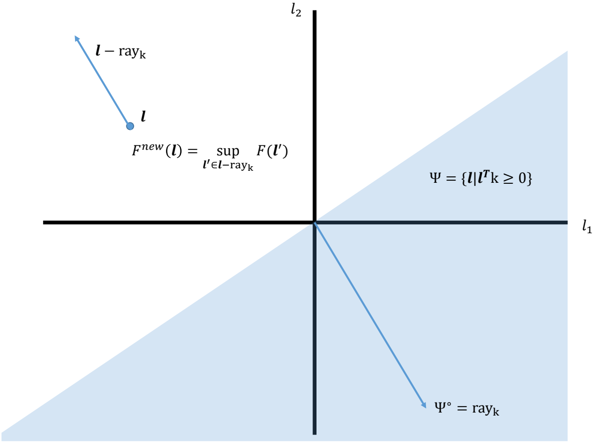


Figure: Sketch map for the value of $F^{new}(l)$ in 2 dimension

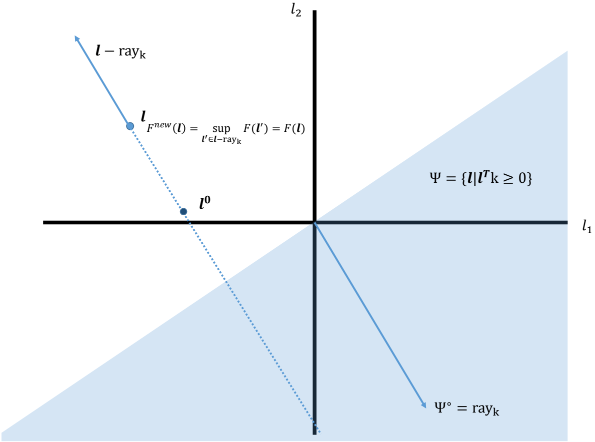


Figure: Sketch map for the value of $F^{new}(l)$ in 2 dimension

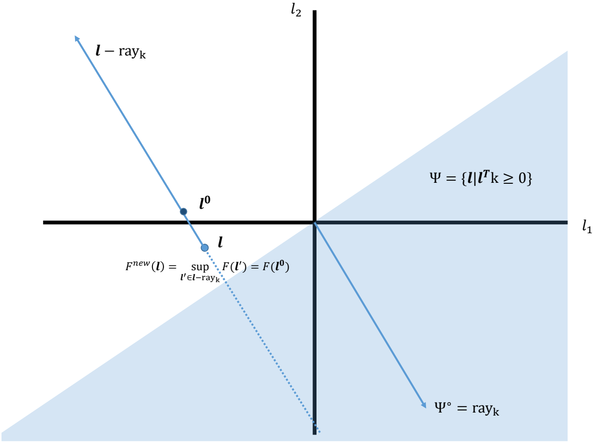


Figure: Sketch map for the value of $F^{new}(l)$ in 2 dimension

Critical Region

$$F^{new}(l) = \sup_{\beta \geq 0 | \; l^1 = l - \beta k} F(l^1) = \begin{cases} F(l) & l^0(l) \notin l - ray_k \\ F(l^0(l))(\geq F(l)) & l^0(l) \in l - ray_k \end{cases}$$

The most important thing for determining F^{new} is to determine the critical region $l^0(R^n)$.

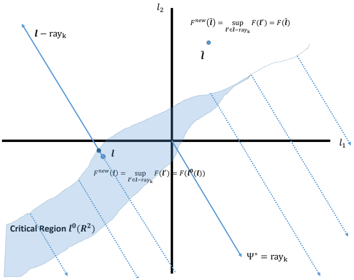


Figure: Sketch map of the critical region

The most important thing for determining F^{new} is to determine the critical region $l^0(R^n)$.

$$l \in l^0(R^n) \Longleftrightarrow 0 = \nabla_\beta F(l - \beta k)|_{\beta=0} = -\nabla F(l)^T k \Longleftrightarrow \nabla F(l) \in k^\perp \Longleftrightarrow l \in \partial g(k^\perp)$$

So finally, it yields

$$F^{new}(l) = \begin{cases} F(l) & l \in \partial g(k^\perp) - ray_k \\ F(l^0(l))(\geq F(l)) & l \in \partial g(k^\perp) + ray_k \end{cases}$$

Notice $int(dom\, f \cap [0,1]^n) = int(dom\ g) \subset dom\ \partial g \subset dom\ g = dom\, f \cap [0,1]^n$

Example

For instance, we set $f(v) = -g(v) = -\log(v_1) - \log(v_2)$ $v = (v_1, v_2) \in (0,1]^2$. Notice that $f(v)$ is strictly convex, thus $F(l)$ will be differentiable.

Since in this case the function $f(v)$ can be separated into the sum of $f_1(v_1)$ and $f_2(v_2)$, we can easily calculate $F(l)$ and $v(l)$.

$$l_1(v_1) = \partial g_1(v_1) = \begin{cases} 1/v_1 & v_1 \in (0,1) \\ (-\infty, 1] & v_1 = 1 \end{cases}$$

Thus, $v_1(l_1) = (\partial g_1)^{-1}(l_1) = \begin{cases} 1/l_1 & l_1 \in (1, +\infty) \\ 1 & l_1 \in (-\infty, 1] \end{cases}$

Then $F_1(l_1) = v(l_1)l_1 - g_1(v_1(l_1)) = \begin{cases} 1 + \log l_1 & l_1 \in (1, +\infty) \\ l_1 & l_1 \in (-\infty, 1] \end{cases}$

Similarly, we can obtain $F_2(l_2)$. It yields $F(l) = F_1(l_1) + F_2(l_2)$.

Suppose the curriculum region $\Psi = \{v|v^T k \geq 0\}$ satisfies the previous conditions in section 1. Besides, $0 \leq k_1 = 1 \leq -k_2$.

Then the critical region with respect to Ψ,

$$\partial g(k^{\perp} \cap dom\, f \cap [0,1]^n) = (1,+\infty)\begin{pmatrix} 1 \\ -k_2 \end{pmatrix} \cup \begin{pmatrix} (-\infty, 1] \\ -k_2 \end{pmatrix}.$$

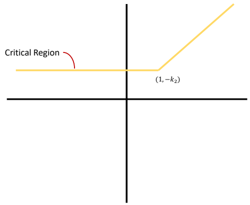
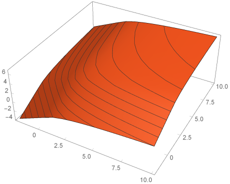
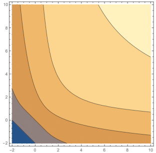


Figure: Critical Region

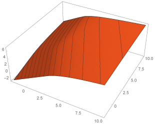
$$F^{new}(l) = \begin{cases} F(l) & l_2 > -k_2 l_1 \;\; and \;\; l_2 > -k_2 l_1 \\ F(l_1 + \frac{l_2 + k_2 l_1}{-2k_2}, -k_2 l_1 + \frac{l_2 + k_2 l_1}{2}) & (l_2 \le -k_2 \;\; or \;\; l_2 \le -k_2 l_1) \; and \; l_2 \ge k_2(l_1 - 2) \\ F(l_1 - \frac{k_2 + l_2}{k_2}, -k_2) & (l_2 \le -k_2 \;\; or \;\; l_2 \le -k_2 l_1) \; and \; l_2 < k_2(l_1 - 2) \end{cases}$$



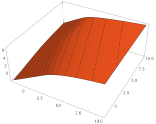
(a)$F(l)$ 3D



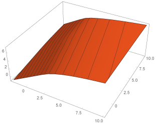
(b) $F(l)$ contour



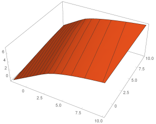
(a) k=-2



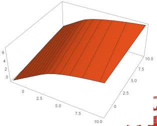
(b) k=-3



(c) k=-4



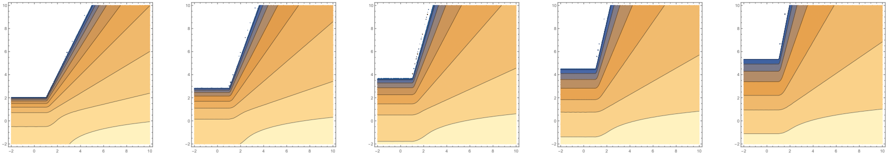
(b) k=-5



(c)k=-6

The following picture illustrates the influence of

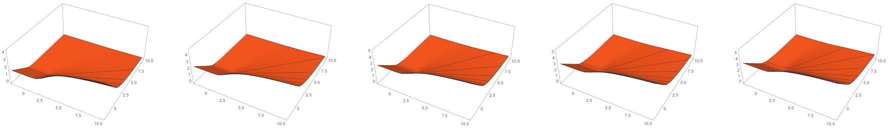
$$\Psi = \{v | v^T \begin{pmatrix} 1 \\ k_2 \end{pmatrix} \geq 0 \, for \; k_2 = -2, \cdots, -6$$



(a) k=-2 (b) k=-3 (c) k=-4 (b) k=-5 (c)k=-6

Figure: $\log[F^{new}(l) - F(l)]$

$F^{new}(l) - F(l) = 0$ on the white area.



(a) k=-2 (b) k=-3 (c) k=-4 (b) k=-5 (c)k=-6

Figure: $F^{new}(l) - F(l)$

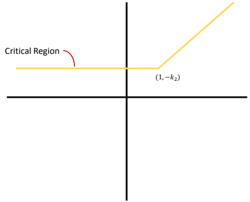


Figure: Critical Region

Conjecture

Can the Critical Region always be calculated by $\partial g(e)$ where e is $boudary(\Psi) \cap dom\ g$?

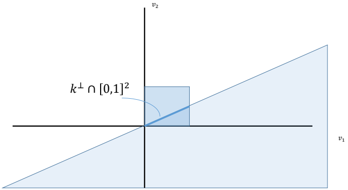


Figure: Critical Edge

Summary

Our work

- Theoretically improves the Self-paced learning
- Provides with two general approaches to design the SPL model
- Theoretically improves the Self-paced curriculum learning
- Illustrates the influence of the some curriculum
- may be applied to Non-convex analysis

For Further Reading I

R.Tyrrell Rockafellar.
Convex Analysis.
PRINCETON UNIVERSITY PRESS, 1997.

Deyu Meng, Qian Zhao.
What’ s The Insight of Self-paced Learning

- Lu Jiang, Deyu Meng, Qian Zhao
  Self-paced Curriculum Learning
  AAAI

THANK YOU

Thank You!