積分型の損失関数を用いたパラメータ推定

Akifumi Okuno[1,2]

[1]ISM, [2]RIKEN AIP

(1) Akifumi Okuno. Minimizing robust density power-based divergences for general parametric density models. arXiv:2307.05251 revised and under review.

(2) Akifumi Okuno. A stochastic optimization approach to train non-linear neural networks with a higher-order variation regularization. arXiv:2308.02293 in preparation for resubmission.

自己紹介

- 阪大基礎工(学部) ⇒ 同(修士) ⇒ 京都大学情報学研究科(博士)
- 指導教員: 下平英寿教授

所属ばかりが増える

(1) 統計数理研究所　数理推論研究系　助教
(2) 同・統計思考院　助教
(3) 同・統計的機械学習センター　助教
(4) 総合研究大学院大学　助教
(5) 理化学研究所　AIPセンター　客員研究員

YouTubeチャンネルをやっています: `https://www.youtube.com/@oao6215`

Overview

(1) robust density-power divergence (arXiv:2307.05251):

$$-\frac{1}{\beta}\frac{1}{n}\sum_{i=1}^{n} p_\theta(x_i)^\beta + \underbrace{\frac{1}{1+\beta}\int p_\theta(x)^{1+\beta}\mathrm{d}x}_{\text{bias correction}},$$

(2) higher-order variation regularization (arXiv:2308.02293):

$$-\frac{1}{n}\sum_{i=1}^{n}\{y_i - f_\theta(x_i)\}^2 + \eta \underbrace{\int_\Omega \left|\frac{\partial^k}{\partial x^k} f_\theta(x)\right|^q \mathrm{d}x}_{\text{variation regularization}}.$$

- Previous works consider (i) numerical integration, or (ii) restricted models.
- We simply apply Robbins and Monro (1951) to minimize the above loss functions.

Robust divergence minimization
(arXiv:2307.05251, revised and under review)

Kullback-Leibler Minimization $\Leftrightarrow$ Likelihood Maximization

- Observations: $x_1, x_2, \ldots, x_n \sim Q$
- We estimate Q by a probabilistic model P_θ (whose p.d.f. is p_θ).
- $\widehat{Q}(x) := n^{-1} \sum_{i=1}^{n} \mathbb{1}(x_i \leq x)$ denotes an empirical distribution.

Then, minimizer of the Kullback-Leibler (KL) cross-entropy

$$d(\widehat{Q}, P_\theta) = -\int \log p_\theta(x) \mathrm{d}\widehat{Q}(x) = -n^{-1} \sum_{i=1}^{n} \log p_\theta(x_i) =: -L(\theta)$$

is equivalent to the maximum likelihood estimator ($=\arg\max_\theta L(\theta)$).

MLE vs Outliers

For normal density,

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\{x_i - \hat{\mu}\}^2.$$

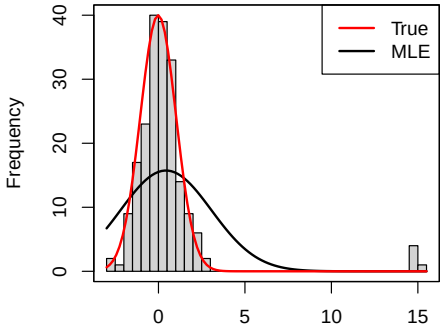


Figure: Outliers adversely affects MLE.

Robust Density-Power Divergence (DPD)

DPD (Basu et al. 1998) $D_\beta(Q,P) = d_\beta(Q,P) - d_\beta(Q,Q)$ is defined with

$$\beta\text{-cross entropy:} \quad d_\beta(\widehat{Q}, P_\theta) = -\frac{1}{\beta} n^{-1} \sum_{i=1}^{n} p_\theta(x_i)^\beta + \frac{1}{1+\beta} \int p_\theta(x)^{1+\beta} \mathrm{d}x.$$

- Typically, power-parameter $\beta = 0.5$ or $\beta = 1$ is employed.
- DPD reduces to KL: $D_\beta \to D$ if $\beta \searrow 0$.
- $\arg\min_P D_\beta(Q,P) = \arg\min_P d_\beta(Q,P) = Q$.

$$\beta\text{-estimator:} \quad \hat{\theta}_\beta := \arg\min_{\theta \in \Theta} d_\beta(\widehat{Q}, P_\theta)$$

is known to be robust against outliers ($\beta > 0$).

Why is DPD robust against outliers?

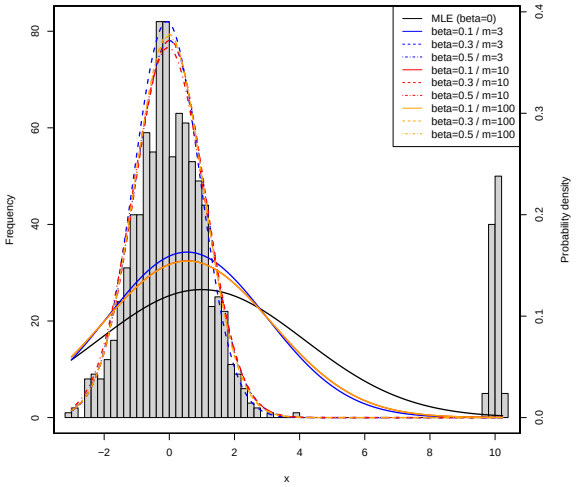
- x is an outlier $\Leftrightarrow p_{\theta_*}(x) \approx 0$.

$$\text{Density power:} \quad \frac{1}{n}\sum_{i=1}^{n} \underbrace{\{-\beta^{-1} p_\theta(x_i)^\beta\}}_{\text{bounded } (\leq 0)} + (\text{bias correction term})$$

$$\text{Kullback Leibler:} \quad \frac{1}{n}\sum_{i=1}^{n} \underbrace{\{-\log p_\theta(x_i)\}}_{\text{unbounded } (\to \infty)}$$

- DPD is upper-bounded while KL is not.

β-estimator vs outliers



Computational Difficulty: How to Minimize Integral-Based Loss?

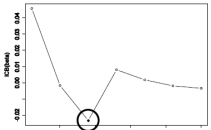
$$d_\beta(\hat{Q}, P_\theta) = -\frac{1}{\beta} n^{-1} \sum_{i=1}^{n} p_\theta(x_i)^\beta + \underbrace{\frac{1}{1+\beta} \int p_\theta(x)^{1+\beta} \mathrm{d}x}_{=: r_\theta^{(\beta)}}.$$

- How to compute the integral term $r_\theta^{(\beta)}$?
- Many studies considers a normal distribution; the term can be calculated as $r_\theta^{(\beta)} = (2\pi\sigma^2)^{-\beta/2}(1+\beta)^{-3/2}$.
- Gradient descent / Newton Raphson, ... is applied.

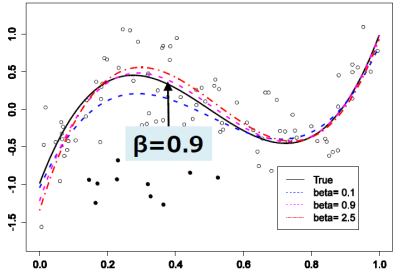
卒論発表 (ちょうど10年前の今頃)

IC_Bを使うと, βを選択できる.

例)3次多項式に限定, IC_Bが最小となるβを選択する.



β=0.9で最小



beta=* : 外れ値を含むデータで, beta=*としてbeta推定.

Figure: ロバスト情報量規準の提案. 誤差項には正規分布を仮定.

What about non-normal densities?

- **Exact**:
  - Exponential (Jones et al., 2001),
  - Generalized-Pareto (Juárez and Schucany, 2004),
  - Weibull (Basu et al., 2016),
  - Generalized exponential (Hazra 2022),
  - Log-normal (rumor...).
- **Approximation**:
  - Gaussian mixture (Fujisawa and Eguchi, 2006) through **upper-bound** minimization,
  - Poisson (Kawashima and Fujisawa, 2019) through **finite approximation**,
  - Skew-normal (Nandy et al. 2022) through **finite approximation**.

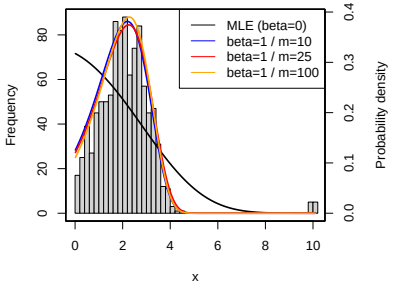
What can we do if x_i follows a remaining distribution... (?)

Contribution of this study

We propose an optimization approach to minimize the DPD for *general parametric density*.

Example: gompertz density $\quad p_\theta(x) = \lambda \exp\left(\omega x + \frac{\lambda}{\omega}\{1 - \exp(\omega x)\}\right), \quad (x \geq 0).$

$$r_\theta^{(\beta)} := \frac{1}{1+\beta} \int p_\theta(x)^{1+\beta} \mathrm{d}x = ??$$



Fullbatch vs stochastic gradient descent (Robbins and Monro, 1951)

To minimize a loss function $A(\theta)$,

- **fullbatch** gradient descent: $\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta \nabla A(\theta^{(t)})$,
- **stochastic** gradient descent: $\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta_t g_t(\theta^{(t)})$,

where

$$\eta, \eta_t > 0,\ \eta_t \searrow 0,\ \text{and}\ \mathbb{E}(g_t(\theta^{(t)})) = \nabla A(\theta^{(t)}).$$

Roughly speaking, we can prove under some assumptions that

$$A(\theta^{(t)}) \to^p \min_{\theta \in \Theta} A(\theta).$$

- We do not need to calculate the exact integral.

Proposal: (unbiased) stochastic gradient for DPD

With $y_1^{(t)}, y_2^{(t)}, \ldots, y_m^{(t)} \sim \tilde{p}$, we define

$$g_t(\theta^{(t)}) = -n^{-1} \sum_{i=1}^{n} p_{\theta^{(t)}}(x_i)^\beta \nabla \log p_{\theta^{(t)}}(x_i) + \frac{1}{m} \sum_{j=1}^{m} \frac{p_{\theta^{(t)}}(y_j^{(t)})}{\tilde{p}(y_j^{(t)})} p_{\theta^{(t)}}(y_j^{(t)})^\beta \nabla \log p_{\theta^{(t)}}(y_j^{(t)}). \quad (1)$$

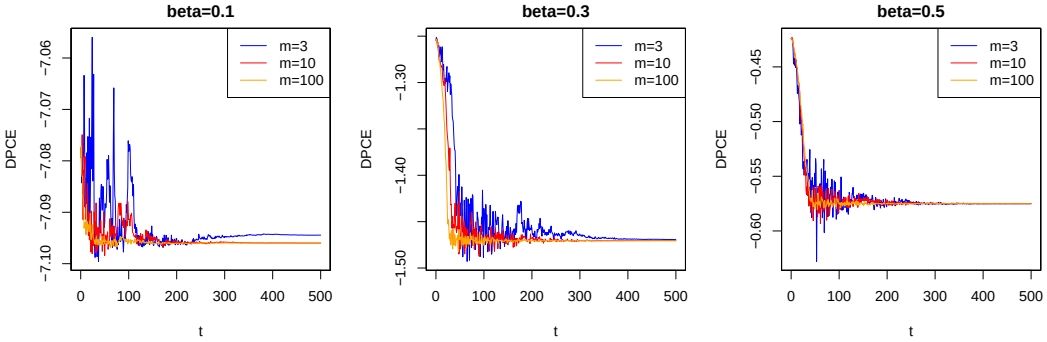
Then, the stochastic gradient is unbiased:

$$\begin{aligned} \mathbb{E}_Y(g_t(\theta^{(t)})) &= -n^{-1} \sum_{i=1}^{n} p_{\theta^{(t)}}(x_i)^\beta \nabla \log p_{\theta^{(t)}}(x_i) + \int p_{\theta^{(t)}}(x)^{1+\beta} \nabla \log p_{\theta^{(t)}}(x) \mathrm{d}x \\ &= \nabla d_\beta(\widehat{Q}, P_{\theta^{(t)}}), \quad \text{(for arbitrary } m \in \mathbb{N}\text{)}. \end{aligned}$$

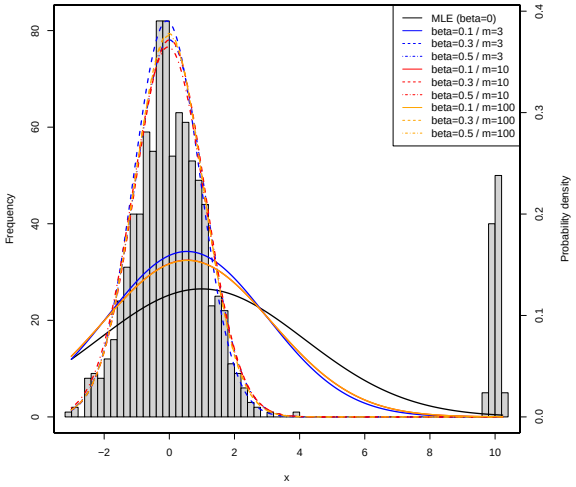
- Similar approach can be found in contrastive divergence (Hinton et al. 2002).

Illustration

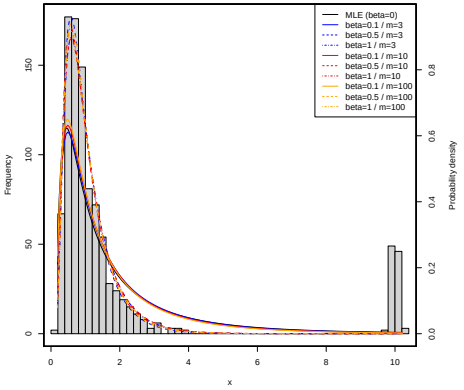
We can monitor the explicit DPD for normal distribution: $(n = 1000, \xi = 0.1)$



Normal density, $\xi = 0.1$.



Inverse Gaussian density, $\xi = 0.1$.



$$p_\theta(x) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right)$$

(*Explicit for of $r_\theta^{(\beta)}$ cannot be obtained for Inverse Gaussian.)

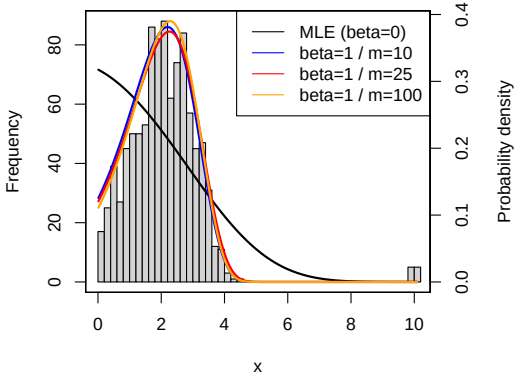


Figure: Gompertz

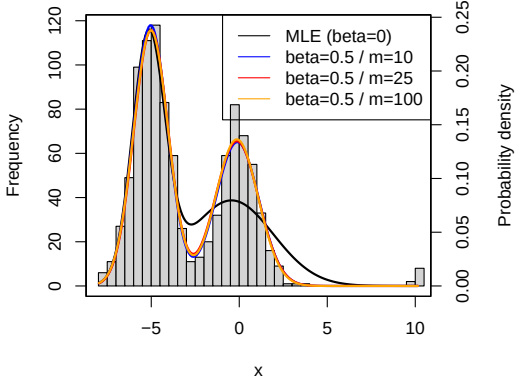
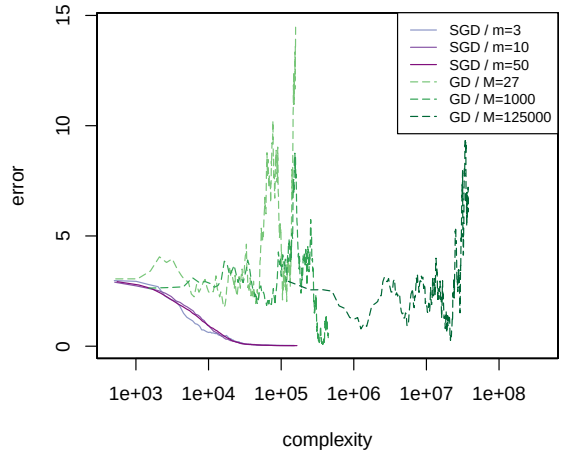


Figure: Gaussian mixture

Multivariate normal mean estimation ($d = 3$).

- SGD complexity: $t(n + m)$,
- GD+numer. int.: $t(n + M)$.



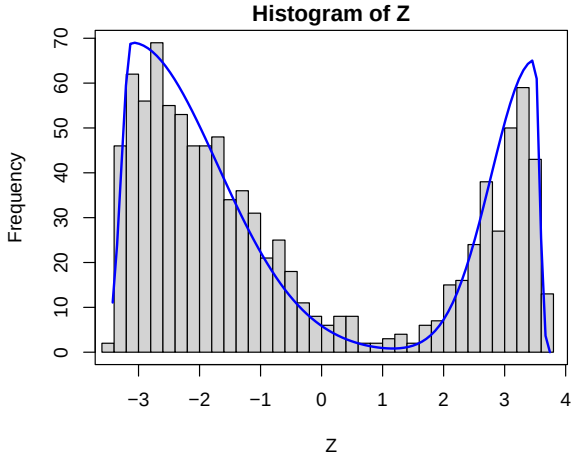

```
install.packages(
    "https://okuno.net/R-packages/sgdpd_1.0.0.zip",
    repos=NULL, type="win.binary")
```

```
sgdpd(f=f, Z=Z, lr=0.1, theta0=c(0,2), exponent=0.2)
```

- https://github.com/oknakfm/sgdpd

```
# model specification (skew normal mixture density)
f_snm <- function(z, theta){
    alpha = sigmoid(theta[7])
    p1 = dsnorm(x=z, mean=theta[1], sd=theta[2], xi=theta[3])
    p2 = dsnorm(x=z, mean=theta[4], sd=theta[5], xi=theta[6])
    alpha * p1 + (1-alpha) * p2
}

# parameter estimation
par_snm = sgdpd(f=f_snm, Z=Z, lr=0.1,
    theta0=c(-1,1,1,1,1,1,0),
    positives=c(2,3,5,6), exponent=0.2)
```



- Density estimation / regression / classification with i.i.d. assumption.

Summary so far

`https://arxiv.org/abs/2307.05251`

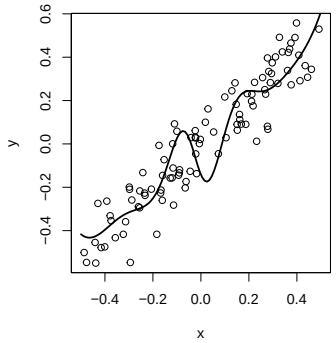
- Historically, normal density ($+\alpha$) has been employed for robust-divergence.
- This study applies stochastic optimization to DPD for general models.
- SGD has been studied for more than 70 years (see, e.g., Robbins and Monro, 1951).
- SGD vs GD + numerical integration: see, e.g., Nemirovski et al. (2009).
- Stochastic approach is compatible with robust estimation (non-convex optimization).
- A Similar approach can be found in contrastive divergence (Hinton et al. 2002).
- γ-divergence (Fujisawa and Eguchi, 2008) can be minimized in the similar way.
- `https://github.com/oknakfm/sgdpd`

We would like to thank A. Nitanda, H. Fujisawa, S. Hashimoto, T. Kawashima(s), K. Harada, K. Yano for helpful comments.

Higher-order variation regularization
(arXiv:2308.02293, in preparation for resubmission)

Motivation

- Nowadays, people use many non-linear models (neural networks, generalized additive models, ...)
- Highly-expressive non-linear models may
  (1) overfit to the dataset,
  (2) fall into a local minima, ...



- We want to obtain a “simpler” curve.

Higher-order variation regularization (HOVR)

Assume the smoothness on $f : \Omega \to \mathbb{R}$, and define (k, q)-th variation regularization:

$$C_{k,q}(f) := \int_\Omega |f^{[k]}(x)|^q \mathrm{d}x, \quad f^{[k]}(x) = \frac{\partial^k f(x)}{\partial x^k}.$$

- Small (k, q)-VR directly yields simpler f.

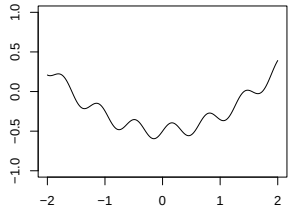


Figure: $f^{[2]}$ is large: $C_{2,2}(f) \approx 197$.

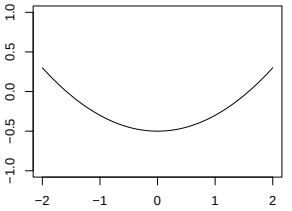


Figure: $f^{[2]}$ is small: $C_{2,2}(f) \approx 0.64$

- (1,1)-VR is known as a *total variation* regularization.

We consider a loss function using (k, q)-VR:

$$L_\eta(\theta) := n^{-1} \sum_{i=1}^{n} \{y_i - f_\theta(x_i)\}^2 + \eta \underbrace{\int_\Omega \left| \frac{\partial^k f_\theta(x)}{\partial x^k} \right|^q \mathrm{d}x}_{\text{HOVR}}.$$

- We may compute SGD with the (unbiased) stochastic gradient:

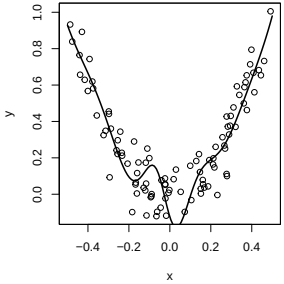
$$g_t(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \nabla \{\tilde{y}_i - f_\theta(\tilde{x}_i)\}^2 + \eta \frac{1}{M} \sum_{j=1}^{M} \nabla |f_\theta^{[k]}(z_j)|^q \quad z_j \sim U(\Omega).$$

Then, under some assumptions, we have

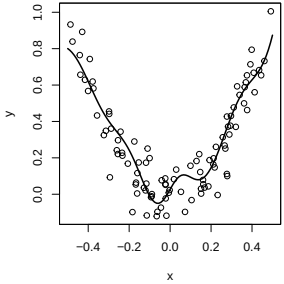
$$L_\eta(\theta^{(t)}) \to \min_{\theta \in \Theta} L_\eta(\theta).$$

Demonstration

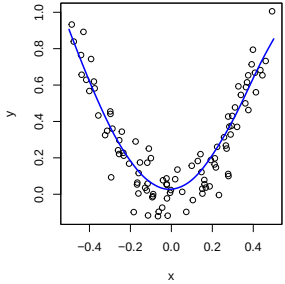
- 1-hidden-layer perceptron with $L = 50$ hidden units and **tanh** activation.
- Same optimizer, same setting, except for the regularization.



(a) No regularization

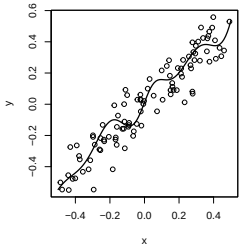


(b) L_2 regularization

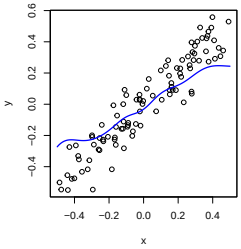


(c) **Proposal**: $(3, 2)$-VR

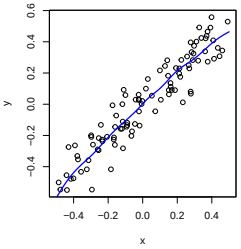
Experiments: linear



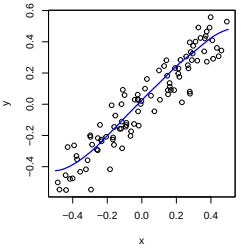
(a) L_2 regularization



(b) $k = 1$-variation reg.

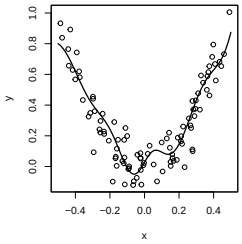


(c) $k = 2$-variation reg.

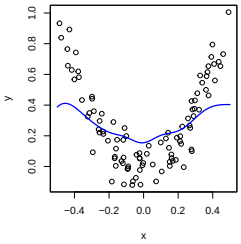


(d) $k = 3$-variation reg.

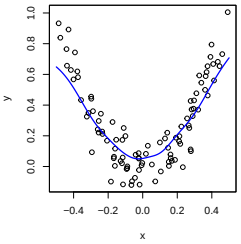
Experiments: quadratic



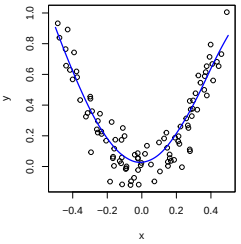
(a) L_2 regularization



(b) $k = 1$-variation reg.

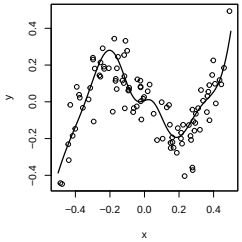


(c) $k = 2$-variation reg.

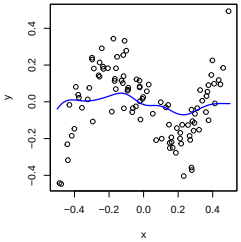


(d) $k = 3$-variation reg.

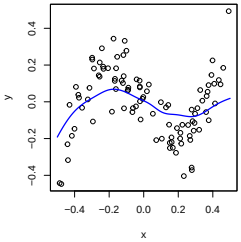
Experiments: cubic



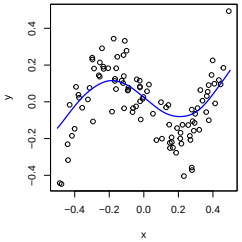
(a) L_2 regularization



(b) $k = 1$-variation reg.

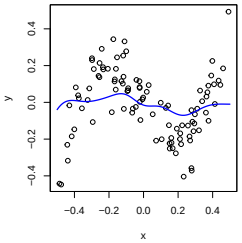


(c) $k = 2$-variation reg.

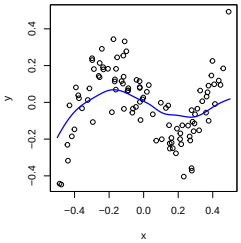


(d) $k = 3$-variation reg.

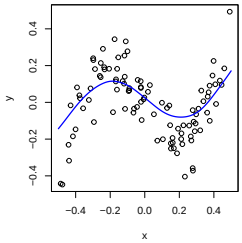
Which variation order should be regularized?



(a) $k = 1$-variation reg.



(b) $k = 2$-variation reg.



(c) $k = 3$-variation reg.

- $k = 1$: piece-wise constant
- $k = 2$: piece-wise linear
- $k = 3$: ??? (seems the best for me, in terms of the "simplicity")

Small k-th variation $\Rightarrow$ small k'-th variation ($k' \leq k$, Sobolev's inequality.)

Summary so far

https://arxiv.org/abs/2308.02293

- We applied SGD to minimize the regression loss function equipped with the higher-order variation regularization (HOVR).
- Compared to the spline regression, we can easily implement the stochastic optimization.
- Stochastic algorithm can be simply generalized to different problems (i.e., classification).
- Also we can simply generalize this approach to multivariate case.
- While previous studies consider penalizing lower-order derivative (mainly, $k = 1$), penalizing higher-order derivatives seems better.
- https://github.com/oknakfm/HOVR