A stochastic optimization approach to minimize robust density power-based divergences for general parametric density models (Okuno 2023, arXiv:2307.05251)

Akifumi Okuno[1,2]

[1]ISM, [2]RIKEN AIP

With my sincere respect to Prof. Basu (ISI Kolkata) for
his team's fascinating paper published in Biometrika (1998).

Overview

Density-power cross-entropy appeared in many presentations so far:

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

Due to the computational intractability,

- **Previous studies**: limited to restricted models (Normal, Weibull, ...).
- **This study**: arbitrary models.

Background

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).
- $\widehat{Q}(x) := n^{-1}\sum_{i=1}^n \mathbb{1}(x_i \le x)$ 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

MLE (with a normal model) is sensitive to outliers.

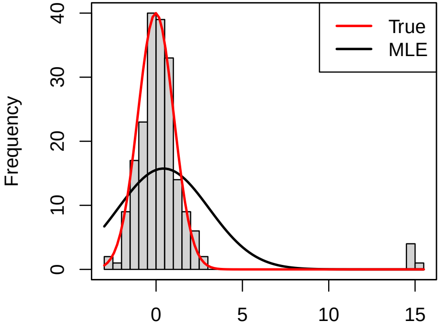


Figure: Outliers adversely affects MLE.

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

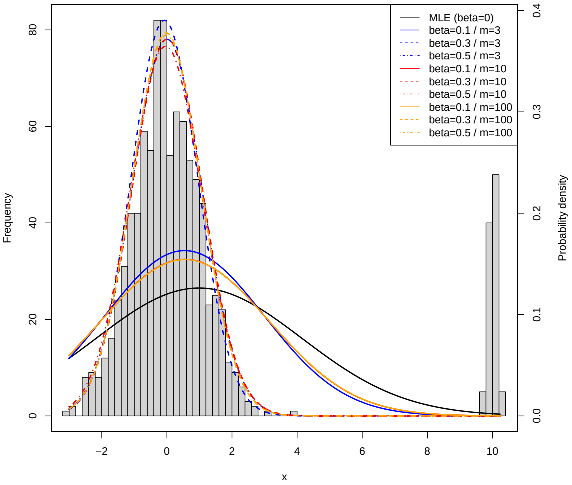
Density-power cross entropy: $$d_\beta(\widehat{Q}, P_\theta) = -\frac{1}{\beta} n^{-1} \sum_{i=1}^{n} p_\theta(x_i)^\beta + \frac{1}{1+\beta} \int p_\theta(x)^{1+\beta} \mathrm{d}x.$$

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

Density-power estimator: $$\hat{\theta}_\beta := \underset{\theta \in \Theta}{\arg\min}\, d_\beta(\widehat{Q}, P_\theta)$$

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

DP-estimator vs outliers

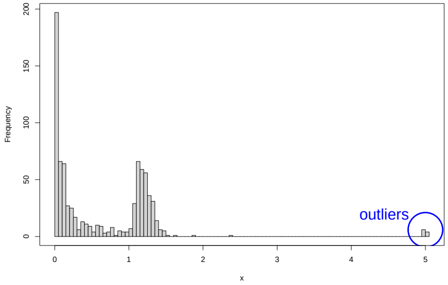


Computational Difficulty: How to Minimize Integral-Based Loss?

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

- How to compute the integral term?
- 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 to obtain density-power estimator.

What can we do if $\{x_i\}$ follow a non-normal distribution?



- Observed vectors seem to follow non-normal distribution.
- Should we use normal models even in this setting?
  $\Rightarrow$ Inevitable model misspecification contradicts to the concept of "robust" estimation.
- General optimization method is greatly appreciated.

DPD minimization for non-normal densities

Few studies employ non-normal models. A list of works I know:

- **Exact** (integral term can be expanded analytically):
  - Exponential (Jones et al., 2001),
  - Generalized-Pareto (Juárez and Schucany, 2004),
  - Weibull (Basu et al., 2016),
  - Generalized-Exponential (Hazra 2022, preprint, *not exponential family)
  - Log-normal (wind rumor, ongoing).

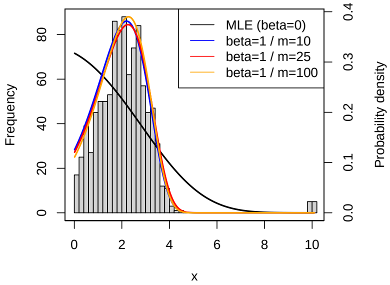
- **Approximation**:
  - Gaussian mixture (Fujisawa and Eguchi, 2006) through **upper-bound** minimization,
  - Poisson (Kawashima and Fujisawa, 2019) through **finite approximation**.
  - Skew-normal (Nandy et al., 2021) through **finite approximation**.

Contribution of this study

We provide an optimization method to minimize the DPD for *general parametric density*.

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

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



- Even mixtures of intricate densities can be optimized!

Proposal

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

$$\theta^{(t)} \to^p \operatorname*{arg\,min}_{\theta \in \Theta} A(\theta).$$

- While **exact integral is needed** to compute **full-batch gradient** $\nabla A(\theta^{(t)})$,
- we can define a **stochastic gradient** $g_t(\theta^{(t)})$ **without integral**!

Proposal: (unbiased) stochastic gradient for DPD

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

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

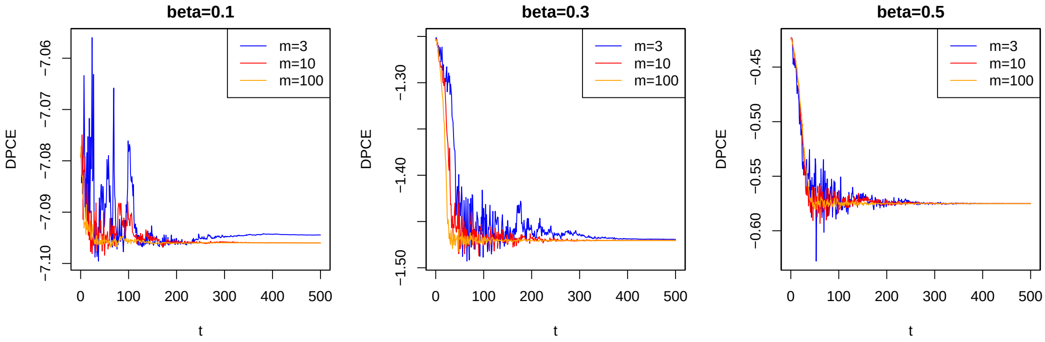
Then, the stochastic gradient is unbiased:

$$\mathbb{E}_Y(g_t(\theta^{(t)})) = \nabla d_\beta(\widehat{Q}, P_{\theta^{(t)}}), \quad \text{(for arbitrary } m \in \mathbb{N}).$$

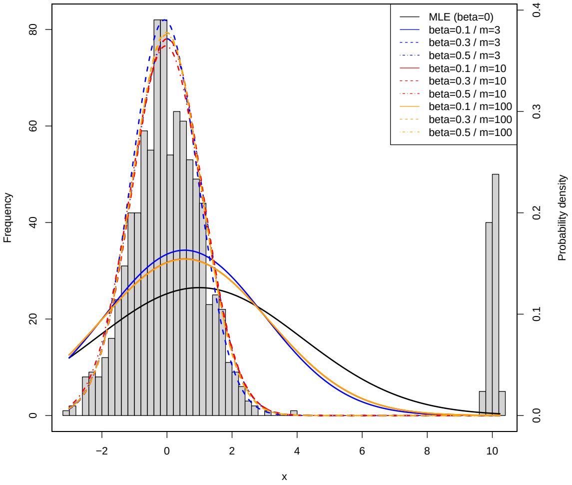
- SGD $\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta_t g_t(\theta^{(t)})$ yields DP-estimator.
- Theoretically, even $m = 1$ is enough to obtain the exact estimator.
- A similar approach can be found in contrastive divergence (Hinton et al. 2002).

Illustration

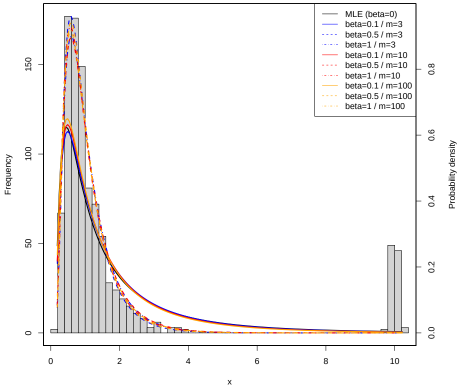
We can monitor the explicit DP-cross entropy for normal distribution: $(n = 1000, \xi = 0.1)$



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.)

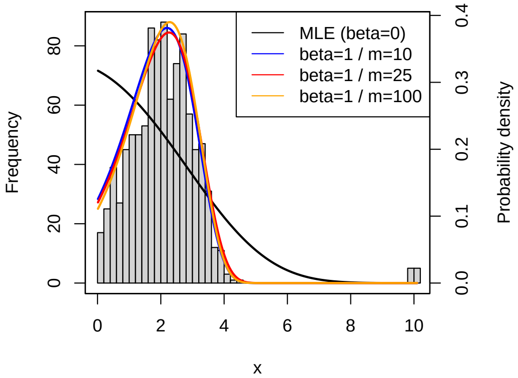


Figure: Gompertz

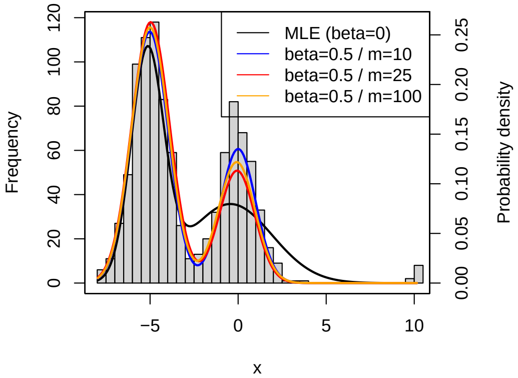


Figure: Gaussian mixture

SGD vs Fullbatch GD (+Numerical Integration)

Comparison with GD (+Numerical integration) with M lattice points.

- **Error**: $O_p(1/\sqrt{T})$ for SGD $< O(1/\sqrt{T} + 1/\sqrt{M})$ for GD+NI
- **Computational efficiency**: $n + m$ for SGD $\ll n + M$ for NI, for each step

See, e.g., Nemirovski et al. (2009) for more detailed comparisons.

- **Robust loss is non-convex**: as is well-known in deep learning theories, stochastic approaches are highly compatible with non-convex loss as SGD is expected to escape from local minima.
- Empirically speaking, for $d > 3$, **SGD is stable** while GD+NI sometimes diverges.

Overall, SGD is highly compatible with the robust divergence minimization.

Conclusion

This presentation was based on `https://arxiv.org/abs/2307.05251`

- We applied a stochastic optimization to DPD for general models.
- SGD has been studied for more than 70 years (see, e.g., Robbins and Monro, 1951).
- Similar approach can be found in Contrastive-divergence (Hinton et al. 2002).
- γ-divergence (Fujisawa and Eguchi, 2008) can be minimized similarly.

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

Please feel free to contact me: A. Okuno (okuno@ism.ac.jp)

Appendix

Why is DPD robust against outliers?

- x is an outlier $\Leftrightarrow p_{\theta_*}(x) \approx 0$.
- DPD is upper-bounded while KL is not.

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

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

γ-divergence minimization

Kanamori and Fujisawa (2015) proved the identity with the γ-divergence (Fujisawa and Eguchi, 2008):

$$\arg\min_{\theta \in \Theta} d_\gamma(\hat{Q}, P_\theta) = \arg\min_{\theta \in \Theta} \left\{ \min_{c>0} d_\beta(\hat{Q}, c \cdot P_\theta) \right\},$$

where $c \cdot P_\theta$ is called unnormalized models.

Therefore, minimization of

$$d_\beta(\hat{Q}, c \cdot P_\theta)$$

with respect to $\psi = (c, \theta)$ yields γ-estimator $\hat{\theta}_\gamma$, which minimizes $d_\gamma(\hat{Q}, P_\theta)$.