Bounds with Data and an Almost Sure Central Limit Theorem using Stein's Method

Gesine Reinert

Department of Statistics
University of Oxford
reinert@stats.ox.ac.uk

Workshop on New Directions in Stein's Method
IMS Singapore, 28 May 2015

Outline

1. What about the data?
2. Maximum likelihood estimators and confidence intervals
3. The effect of the prior on the posterior in Bayesian analysis
4. An almost sure central limit theorem
5. Summary and Outlook

Collaborations with Andreas Anastasiou, George Deligiannidis, Larry Goldstein, Christophe Ley and Yvik Swan

Stein's method - what about the data?

Stein's method is used to obtain bounds in distributional distances.

A motivation for these distances is that real data sets are always finite and hence asymptotic results should be quantified.

Usually in Stein's method the input are random variables.

Uusally in statistics the input are observations.

How to bridge this gap?

Three examples

Here we look at three examples.

1. Maximum likelihood estimators and confidence intervals.
2. The effect of the prior on the posterior in Bayesian analysis.
3. An almost sure central limit theorem.

Distances

We use the bounded Wasserstein distance, with random variables standing for their distributions,

$$d_{bW}(F, G) = \sup \{|\mathbb{E}[h(F)] - \mathbb{E}[h(G)]| : h \in H\},$$

with

$$H = \left\{ h : \mathbb{R} \to \mathbb{R} \ : \sup_{\substack{x \neq y \\ x,y \in \mathbb{R}}} \frac{|h(x) - h(y)|}{|x - y|} + \|h\| \leq 1 \right\}.$$

We also use the Wasserstein distance,

$$d_{W}(F, G) = \sup \{|\mathbb{E}[h(F)] - \mathbb{E}[h(G)]| : h \in H\},$$

with

$$H = \left\{ h : \mathbb{R} \to \mathbb{R} \ : \sup_{\substack{x \neq y \\ x,y \in \mathbb{R}}} \frac{|h(x) - h(y)|}{|x - y|} \leq 1 \right\}.$$

Maximum likelihood estimators and confidence intervals

(joint work with Andreas Anastasiou, and discussions with Robert Gaunt)

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be i i.i.d. with joint density function $f(\boldsymbol{x}|\theta)$, with unknown parameter $\theta \in \Theta \subset \mathbb{R}$. Let θ_0 be the true parameter.

For observations $\mathbf{x} = (x_1, \ldots, x_n)$ estimate θ by $\hat{\theta} = \hat{\theta}_n(\boldsymbol{x})$ which maximises the likelihood $L(\theta; \boldsymbol{x}) = f(\boldsymbol{x}|\theta)$.

We assume that $\hat{\theta}$ exists and is unique; that $l = \log L$ is smooth both in x and in θ, that $\mathbb{E}_\theta[l'(\theta; \boldsymbol{X})] = 0$.

Define the Fisher information $i(\theta)$ through

$$\mathrm{Var}_\theta[l'(\theta; \boldsymbol{X})] = \mathbb{E}_\theta(-l''(\theta_0; \boldsymbol{X})) = n\, i(\theta)$$

and assume that $i(\theta_0) \neq 0$.

Theorem

(Fisher, 1925)
Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with probability density (or mass) function $f(x_i|\theta)$, where θ is the scalar parameter. Assume that the MLE exists and it is unique and some regularity conditions are satisfied. Then

$$(a) \qquad \frac{1}{\sqrt{n}} l'(\theta_0; \boldsymbol{X}) \xrightarrow[n\to\infty]{\mathrm{d}} \mathrm{N}(0, i(\theta_0))$$

$$(b) \qquad \sqrt{n\, i(\theta_0)}(\hat{\theta}_n(\boldsymbol{X}) - \theta_0) \xrightarrow[n\to\infty]{\mathrm{d}} \mathrm{N}(0, 1).$$

This theorem gives only a qualitative result as $n \to \infty$.

Confidence intervals

This confidence interval takes the observations as input.

For $Z \sim \mathrm{N}(0,1)$, suppose that

$$d_{bW}\left(\sqrt{n\, i(\theta_0)}(\hat{\theta}_n(\boldsymbol{X}) - \theta_0), Z\right) = B_{bW}.$$

If $i(\theta_0)$, is known, then a conservative $100(1-\alpha)\%$ confidence interval for θ_0 is given by

$$\left(\hat{\theta}_n(\boldsymbol{x}) - \frac{\Phi^{-1}\left(1 - \frac{\alpha}{2} + 2\sqrt{B_{bW}}\right)}{\sqrt{n\, i(\theta_0)}}, \hat{\theta}_n(\boldsymbol{x}) - \frac{\Phi^{-1}\left(\frac{\alpha}{2} - 2\sqrt{B_{bW}}\right)}{\sqrt{n\, i(\theta_0)}}\right).$$

For applications, B_{bW} should be small.

Heuristic for a normal approximation

We have that $l'(\hat{\theta}_n(\boldsymbol{x}); \boldsymbol{x}) = 0$. Taylor expansion about θ_0 gives

$$l''(\theta_0; \boldsymbol{x}) \left(\hat{\theta}_n(\boldsymbol{x}) - \theta_0\right) = -l'(\theta_0; \boldsymbol{x}) - R_1(\theta_0; \boldsymbol{x}),$$

so that

$$\begin{aligned} -n\, i(\theta_0) \left(\hat{\theta}_n(\boldsymbol{x}) - \theta_0\right) &= -l'(\theta_0; \boldsymbol{x}) - R_1(\theta_0; \boldsymbol{x}) \\ &\quad - \left(\hat{\theta}_n(\boldsymbol{x}) - \theta_0\right) \left[l''(\theta_0; \boldsymbol{x}) + n\, i(\theta_0)\right]. \end{aligned}$$

Re-arranging,

$$\hat{\theta}_n(\boldsymbol{x}) - \theta_0 = \frac{l'(\theta_0; \boldsymbol{x}) + R_1(\theta_0; \boldsymbol{x}) + R_2(\theta_0, \boldsymbol{x})}{n\, i(\theta_0)}.$$

Here

$$\begin{aligned} R_1(\theta_0; \boldsymbol{x}) &= \frac{1}{2} \left(\hat{\theta}_n(\boldsymbol{x}) - \theta_0\right)^2 l^{(3)}(\theta^*; \boldsymbol{x}), \quad \text{some } \theta^*, \text{ and} \\ R_2(\theta_0, \boldsymbol{x}) &= (\hat{\theta}_n(\boldsymbol{x}) - \theta_0) \left(l''(\theta_0; \boldsymbol{x}) + n\, i(\theta_0)\right). \end{aligned}$$

Moreover

$$l'(\theta_0; \boldsymbol{X}) = \sum_{i=1}^{n} \frac{d}{d\theta} \log f(X_i|\theta)$$

is the sum of i.i.d. random variables, and we can apply standard Stein results to this term, such as

Lemma

Let $Y_1, Y_2, \ldots, Y_n$ be independent with $\mathbb{E}(Y_i) = 0, \mathrm{Var}(Y_i) = \sigma^2 > 0$ and $\mathbb{E}|Y_i|^3 < \infty$. Let $W = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$ and $K \sim \mathrm{N}(0, \sigma^2)$. Then

$$d_{bW}(W, K) \leq \frac{1}{\sqrt{n}} \left(2 + \frac{1}{\sigma^3} \left[\mathbb{E}|Y_1|^3 \right] \right).$$

Theorem

Assume further that $\mathbb{E}\left|\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_1|\theta_0)\right|^3 < \infty$ *and* $\mathbb{E}\left(\hat{\theta}_n(\boldsymbol{X}) - \theta_0\right)^4 < \infty$. *Let* $0 < \epsilon$ *be such that* $(\theta_0 - \epsilon, \theta_0 + \epsilon) \subset \Theta$ *and* $Z \sim \mathrm{N}(0,1)$. *Then*

$$
\begin{aligned}
& d_{bW}\left(\sqrt{n\, i(\theta_0)}(\hat{\theta}_n(\boldsymbol{X}) - \theta_0), Z\right) \\
& \leq \frac{1}{\sqrt{n}}\left(2 + \frac{1}{[i(\theta_0)]^{\frac{3}{2}}}\left[\mathbb{E}\left|\frac{\mathrm{d}}{\mathrm{d}\theta}\log f(X_1|\theta_0)\right|^3\right]\right) + 2\frac{\mathbb{E}\left(\hat{\theta}_n(\boldsymbol{X}) - \theta_0\right)^2}{\epsilon^2} \\
& \quad + \frac{1}{\sqrt{n\, i(\theta_0)}}\left\{\mathbb{E}\left(|R_2(\theta_0; \boldsymbol{X})|\,\middle|\, |\hat{\theta}_n(\boldsymbol{X}) - \theta_0| \leq \epsilon\right)\right. \\
& \quad \left. + \frac{1}{2}\left[\mathbb{E}\left(\sup_{\theta:|\theta-\theta_0|\leq\epsilon}\left|I^{(3)}(\theta; \boldsymbol{X})\right|^2 \,\middle|\, |\hat{\theta}_n(\boldsymbol{X}) - \theta_0| \leq \epsilon\right)\mathbb{E}\left(\hat{\theta}_n(\boldsymbol{X}) - \theta_0\right)^4\right]^{\frac{1}{2}}\right\}.
\end{aligned}
$$

- We also have a bound which does not depend on an explicit form of $\hat{\theta}_n$; we can bound the mean square error of $\hat{\theta}_n$ using the theorem for a special Lipschitz function.
- When $\hat{\theta}_n$ is on the boundary of the parameter space with positive probability, such as for the Poisson distribution, then we use a perturbation approach - and we acknowledge very helpful discussions with Robert Gaunt.
- The multivariate parameter version is under way.
- The bound depends on the unknown true parameter θ_0. This is plausible but affects the construction of confidence intervals.
- With this bound we obtain conservative confidence intervals which depend on the data explicitly through $\hat{\theta}(\mathbf{x})$.

The effect of the prior on the posterior in Bayesian analysis

(joint work with Christophe Ley and Yvik Swan)

Given realisations $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ of random variables $X_1, \ldots, X_n$ with joint distribution

$$\pi_1(x_1, x_2, \ldots, x_n|\theta),$$

where θ is a realistion of a random variable Θ, we would like to draw inference on Θ.

Before any observation has been made (a priori) we think that Θ has the (prior) distribution p_0. We update our belief on Θ in light of the observations by applying Bayes' formula, so that the posterior density of Θ, given the observations $\mathbf{y}$, is

$$p_2(\theta|\mathbf{x}) = \pi_1(\mathbf{x}|\theta)p_0(\theta) = \kappa_1(\mathbf{x})p_1(\theta, \mathbf{x})p_0(\theta).$$

Here $p_1(\theta, \mathbf{x})$ is a probability density for θ.

The structure of the problem

We are comparing two distributions whose densities p_1 with support $[a_1, b_1]$ and p_2 are of product type, in the sense that $p_2 = \pi_0 p_1$ for a non-negative function p_0. Let $X_1 \sim p_1$ and $X_2 \sim p_2$.

Assume that p_1 and p_2 are absolutely continuous, that π_0 is differentiable and that for all Lipschitz-continuous functions h with $\mathbb{E}h(X_1) < \infty$,

$$\begin{aligned} \lim_{x \to a_1} \pi_0(x) \int_{a_1}^{x} (h(y) - \mathbf{E}[h(X_1)]) p_1(y) dy &= 0 \\ \lim_{x \to b_1} \pi_0(x) \int_{x}^{b_1} (h(y) - \mathbf{E}[h(X_1)]) p_1(y) dy &= 0. \end{aligned}$$

In general, if $\mathbb{E}X_1 = \mu$ exists, then the Stein kernel $\tau_1 : [a_1, b_1] \to \mathbb{R}$ for X_1 is

$$\tau_1(x) = \frac{1}{p(x)} \int_{-\infty}^{x} (\mu - y) p(y) dy.$$

Theorem

The Wasserstein distance between $X_1 \sim p_1$ and $X_2 \sim p_2 = \pi_0 p_1$ satisfies the following inequalities:

$$|\mathbb{E}X_2 - \mathbb{E}X_1| \leq d_{\mathcal{W}}(X_1, X_2) \leq \mathbb{E}\left[\left|(\log \pi_0(X_2))'\right| \tau_1(X_2)\right],$$

where τ_1 is the Stein kernel associated with p_1.

Heuristical explanation

For a random variable X with continuous univariate density p having support $\mathcal{I}$, define $\mathcal{T}_X$ as Stein operator acting on a suitable class of functions $\mathcal{F}(X)$ through

$$\mathcal{T}_X : \mathcal{F}(X) \to L^1(p) : f \mapsto \mathcal{T}_X f = \frac{(fp)'}{p}.$$

Then for Y with support $\mathcal{I}$,

$$\mathbb{E}[\mathcal{T}_X f(Y)] = 0 \text{ for all } f \in \mathcal{F}(X) \iff Y \sim p.$$

Now if $p_2 = \pi_0 p_1$ then

$$\mathcal{T}_2(f) = \mathcal{T}_1(f) + f\frac{\pi_0'}{\pi_0} = \mathcal{T}_1(f) + f(\log \pi_0)'.$$

Hence

$$\mathcal{T}_2(f) - \mathcal{T}_1(f) = f(\log \pi_0)'.$$

Set $g = f/\tau_1$ and use $||g|| \leq ||h'||$ to obtain the theorem.

Bayesian interpretation

We observe data points $x := (x_1, x_2, \ldots, x_n)$ with sampling distribution $\pi_1(x \,|\, \theta)$. We take θ, the one dimensional parameter, to be distributed according to some (possibly improper) prior $p_0(\theta)$, and let the posterior be given by $p_2(\theta; x) \propto p_0(\theta) p_1(\theta; x)$. Set

$$\Theta_1 \sim p_1(\theta; x) = \kappa_1(x)\pi_1(x; \theta)$$

and

$$\Theta_2 \sim p_2(\theta; x) = \kappa_2 \pi_0(\theta)\pi_1(x, \theta).$$

Then our theorem applies,

$$d_{\mathcal{W}}(\Theta_1, \Theta_2) \leq \frac{\kappa_2}{\kappa_1} \mathbb{E} \left| \pi_0'(\Theta_1)\tau_1(\Theta_1) \right|,$$

and we can assess the influence of the prior on the posterior.

Example: Binomial model, Beta prior

Assume $x \sim \mathit{Binomial}(n, \theta)$, with known n, and the prior for θ is

$$\pi_0(\theta) = \kappa_0 \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \theta \in [0,1],$$

with $\alpha > 0$ and $\beta > 0$. Then $\tau_1(\theta) = \frac{\theta(1-\theta)}{n+2}$. A direct computation gives

$$d_{\mathcal{W}}(\Theta_1, \Theta_2) \leq \frac{1}{n+2}\left(|2-\beta-\alpha|\frac{\alpha+x}{\alpha+\beta+n} + |\alpha-1|\right).$$

- Unless $\alpha = 1$ the bound will be of order $1/n$ no matter how favourable x is.
- If $\alpha = 1$ but $\beta \neq 1$ then the bound is smallest when $x = 0$, and is then of order $1/n^2$.
- If $\alpha = 1 = \beta$ then the bound is zero, as it should be as then $p_1 = p_2$, the prior is uniform.

Example: Binomial model, non-informative prior

Using the Haldane prior $p_0(\theta) = \kappa_0(\theta(1-\theta))^{-1}$, direct computation gives

$$d_{\mathcal{W}}(\Theta_1, \Theta_2) \leq \frac{2}{n+2}\left(\left|\frac{x}{n} - \frac{1}{2}\right| + \sqrt{\frac{x(n-x)}{n^2(n+1)}}\right).$$

If $x = \frac{n}{2}$ then the bound is of order $n^{-\frac{3}{2}}$.

Using Jeffreys' prior $p_0(\theta) = \kappa_0(\theta(1-\theta))^{-\frac{1}{2}}$, direct computation gives

$$d_{\mathcal{W}}(\Theta_1, \Theta_2) \leq \frac{1}{n+2}\left(\left|\frac{x+\frac{1}{2}}{n+1} - \frac{1}{2}\right| + \sqrt{\frac{(x+\frac{1}{2})(n-x+\frac{1}{2})}{(n+1)^2(n+2)}}\right).$$

Again if $x = \frac{n}{2}$ then the bound is of order $n^{-\frac{3}{2}}$.

- The bounds appear to be the first explicit bounds of this nature.
- The data appear explicitly in the bounds.
- The multivariate case is under way.

An almost sure central limit theorem

(joint work with George Deligiannidis and Larry Goldstein)

Brosamler (1988) and Schatte (1988) show the following result. For $k \in \mathbb{N}$ and some $\delta > 0$, let

$$S_k = X_1 + \cdots X_k,$$

the k^{th} partial sum of i.i.d. real valued random variables X_i with mean zero, variance 1, and finite $(2+\delta)^{th}$ moment, defined on a probability space $(\Omega, \mathcal{F}, P)$. Then there is a P-null set N such that for all $\omega \in N^c$,

$$\frac{1}{\log n} \sum_{k=1}^{n} \frac{1}{k} \delta_{k^{-\frac{1}{2}} S_k(\omega)} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1) \text{ as } n \to \infty$$

where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution, and $\mathcal{N}(0,1)$ the standard normal distribution. Lacey and Philipp (1990) show that $(\log n)^{-1}$ is the correct scaling in order to get a nontrivial limit.

Can we quantify this result?

Our strategy is as follows. We first consider a deterministic vector $\mathbf{x} = (x_1, \ldots, x_n)$ with distinct values, and consider the empirical (non random) measure

$$\nu_{n,x} = \kappa_n \sum_{k=1}^{n} \frac{1}{k} \delta_{x_k},$$

where $\kappa_n = \left(\sum_{k=1}^{n} \frac{1}{k}\right)^{-1}$.

1. We assess, in terms of $\mathbf{x}$, how far $\nu_{n,x}$ is from a normal distribution.
2. In the next step we show that for $x_k = \frac{S_k}{\sqrt{k}}$ where S_k is the k^{th} partial sum of bounded mean zero variables with variance one, the bound will go to zero almost surely.

Fix $h \in \mathrm{Lip}(1)$ and denote by f the unique bounded solution of the Stein equation

$$h(x) - Nh = f'(x) - xf(x)$$

for the $\mathcal{N}(0,1)$ distribution, where $Nh = Eh(Z)$ for $Z \sim \mathcal{N}(0,1)$.

Theorem

Let $\mathbf{x} = (x_1, \ldots, x_n)$ be a vector of fixed real numbers, not all zero, let

$$\nu_{n,x} = \kappa_n \sum_{k=1}^{n} \frac{1}{k} \delta_{x_k}.$$

Let f denote the unique bounded solution of the Stein equation for h. Then

$$\int h d\nu_{n,x} - Nh = \kappa_n \sum_{k=1}^{n} \frac{1}{k} \left\{ f'(x_k) - x_k f(x_k) \right\}.$$

This equality is true for any $\mathbf{x} = (x_1, \ldots, x_n)$.

Proof.

Let the random index I have distribution

$$\mathbb{P}(I = k) = p_k \text{ where } p_k = \frac{\kappa_n}{k},$$

and let

$$X_I = x_I.$$

Then for any function g,

$$\mathbb{E}[g(X_I)] = \sum_{k=1}^{n} p_k g(x_k) = \int g d\nu_{n,x}.$$

From the Stein equation,

$$\begin{aligned} \int h d\nu_{n,x} - Nh = \mathbb{E}h(X_I) - Nh &= \mathbb{E}f'(X_I) - \mathbb{E}X_I f(X_I) \\ &= \kappa_n \sum_{k=1}^{n} \frac{1}{k} \left\{ f'(x_k) - x_k f(x_k) \right\}, \end{aligned}$$

yielding the assertion.

Why is this helpful?

Corollary

Let

$$\xi_n = \frac{1}{\kappa_n}\sum_{k=1}^{n}\frac{1}{k}\delta_{x_k(\omega)} = \frac{1}{\kappa_n}\sum_{k=1}^{n}\frac{1}{k}\delta_{k^{-\frac{1}{2}}S_k(\omega)},$$

then

$$\begin{aligned}\int h d\xi_n - Nh &= R = R(h)(\omega)\\ &= \kappa_n\sum_{k=1}^{n}\frac{1}{k}\left\{f'\left(k^{-\frac{1}{2}}S_k(\omega)\right) - k^{-\frac{1}{2}}S_k(\omega)f\left(k^{-\frac{1}{2}}S_k(\omega)\right)\right\}.\end{aligned}$$

McDiarmid's concentration inequality and Borel-Cantelli give:

Theorem

Let $h_l \in Lip(C_l)$ *for* $l = 1, 2, \ldots,$ *such that* $\sum_{l=k}^{\infty} \frac{C_l}{l^{\frac{3}{2}}} \leq Ak^{-\frac{1}{2}}$ *for a constant* $A > 0$ *. Let* $y = (y_1, \ldots, y_n)$ *with* $y_i \in [-B, B]$ *and let*

$$g(y) = \kappa_n \sum_{l=1}^{n} \frac{1}{l} \left\{ h_l \left(\frac{1}{\sqrt{l}} \sum_{j=1}^{l} y_j \right) - N h_l \right\}.$$

Let $X = (X_1, X_2, \ldots, X_n)$ *be i.i.d. mean zero, variance 1, with* $|X_i| \leq B$ *and let* $\mu_n(g) = \mathbb{E}g(\mathbf{X})$*. Then for all* $t > 0$

$$\mathbb{P}(|g(\mathbf{X}) - \mu_n(g)| \geq t) \leq 2e^{-\frac{2t^2}{4A^2B^2\kappa_n}}.$$

In particular, $g(\mathbf{X}) - \mu_n(g) \to 0$ *almost surely.*

Now use Stein's method for normal approximation to show that

$$|\mu_n| = \left| \kappa_n \sum_{k=1}^{n} \frac{1}{k} \{\mathbb{E}h(X_k) - Nh\} \right| \leq 3||h'||\kappa_n \left(1 + \mathbf{E}|X_1^3|\right) = O((\log n)^{-1}).$$

Here is the final result.

Theorem

Fix $h \in \mathrm{Lip}(1)$. For all $s > 0$, and $\int h d\xi_n - Nh = R = R(h)(\omega)$,

$$\mathbb{P}(|R| > s) \leq e^{-\frac{2(s-\mu_n)^2}{c}} + e^{-\frac{2(s+\mu_n)^2}{c}}$$

where

$$c = 4A^2B^2\kappa_n.$$

Moreover,

$$R(\omega) \to 0 \quad \textit{almost surely.}$$

Note that

$$d_W\left(\kappa_n \sum_{k=1}^{n} \frac{1}{k} \delta_{k^{-\frac{1}{2}} S_k(\omega)}, \mathcal{N}(0,1)\right) = \sup_{h \in Lip(1)} \left| \int h d\xi_n - Nh \right| = \sup_{h \in Lip(1)} |R(h)(\omega)|$$

but the convergence in the proposition is not (yet) uniform over all h ...

Work is under way!

Some ideas for a uniform result

Let $\mathbf{x} = (x_1, \ldots, x_n)$ be n distinct values.

Derive a Stein operator for the distribution of Y, where

$$\mathbb{P}(Y = x_k) = p_k, \quad k = 1, \ldots, n.$$

using *Ley, Swan and R. (2014)*.

Define

$$\Delta_x f(x_k) = f(x_{k+1}) - f(x_k);$$

the subscript x is a reminder that Δ_x is not the usual forward difference.

The inverse of Δ_x is

$$\Delta_x^{-1} = -\sum_{l=k}^{n} f(x_l).$$

We set as Stein operator $\mathcal{T}$

$$\mathcal{T}f(x_k) = \frac{1}{p(x_k)}\Delta_x(fp)(x_k).$$

Its inverse is

$$\mathcal{T}^{-1}f = \frac{1}{p}\Delta_x^{-1}(fp).$$

Let

$$\mu_n = \mu_n(id) = \sum_{k=1}^{n} p(x_k)x_k.$$

Similarly let σ_n^2 be the variance of Y.

Evaluating $\mathcal{T}^{-1}$ at the function $f = id - \mu_n$ we obtain the so-called Stein kernel τ

$$\tau(x_k) = -\frac{1}{p(x_k)} \sum_{l=k}^{n} (x_l - \mu_n) p(x_l).$$

It follows similarly as for the zero bias construction that

$$-\mathbb{E}\tau(Y)\Delta_x^* f(Y) = \mathbb{E}(Y - \mu_n) f(Y).$$

As the value $x_1, \ldots, x_n$ are not assumed to be ordered, in general τ does not have to be non-negative, and hence does not have to be a density.

Then we have

$$\begin{aligned}
&\mathbb{E}\sigma_n^2 f'(Y) - (Y-\mu_n)f(Y) \\
&\quad = \mathbb{E}\sigma_n^2 f'(Y) + \tau(Y)\Delta_x^* f(Y) \\
&\quad = \sum_k p(x_k) f'(x_k)\sigma_n^2 + \sum_k p(x_k)\tau(x_k)(x_k - x_{k-1})\frac{\Delta_x^* f(x_k)}{x_k - x_{k-1}} \\
&\quad = \sum_k p(x_k)\frac{\Delta_x^* f(x_k)}{x_k - x_{k-1}}\left\{\sigma_n^2 + \tau(x_k)(x_k - x_{k-1})\right\} \\
&\qquad + \sum_k p(x_k)\left(f'(x_k) - \frac{\Delta_x^* f(x_k)}{x_k - x_{k-1}}\right).
\end{aligned}$$

The second term covers the discretisation error. The first term should be boundable!

Possible generalisations:

Almost sure invariance principle (Lacey and Philipp (1990)

Associated sequences, mixing sequences (Peligrad and Shao (1995))

Other averages, independent but not identically distributed (Rychlik and Szuster (2003))

Martingales (Bercu et al. (2009))

Applications to stochastic approximation algorithms (Cenac (2013))

Self-normalised products of partial sums (Wu and Chen (2013))

...

Summary and Outlook

Stein's method can be used to get bounds which depend explictly on the observations.

There are many more statistics problems which could potentially be tackled in this way!