Wasserstein-Cramér-Rao Theory of Unbiased Estimation

Adam Quinn Jaffe

With Nicolás Garciá Trillos and Bodhisattva Sen

Fix a statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\} \subseteq \mathcal{P}(\mathbb{R}^d)$ for $\Theta \subseteq \mathbb{R}$.

For $\theta \in \Theta$, estimate $\chi(\theta)$ from i.i.d. samples $X_1, \ldots, X_n$ from P_θ.

If $T_n : (\mathbb{R}^d)^n \to \mathbb{R}$ is unbiased, then typically try to minimize

$$\mathbb{V}_\theta(T_n) := \mathbb{E}_\theta \left[|T_n(X_1, \ldots, X_n) - \chi(\theta)|^2 \right].$$

The variance is the main focus of most classical statistical theory...

Q: How small can variance be, among all unbiased estimators?

A: Cramér-Rao bound, stating

$$\mathbb{V}_\theta(T_n) \geq \frac{1}{nI(\theta)}$$

for the Fisher information

$$I(\theta) := \mathbb{E}_\theta \left[\left(\dot{\ell}_\theta(X) \right)^2 \right].$$

Q: For which models can this lower bound be exactly achieved?

A: Exponential families,

$$\frac{\mathrm{d}p_\theta}{\mathrm{d}\lambda}(x) = \exp\left((\eta(\theta))^\top \phi(x) - \Lambda(\theta)\right).$$

Q: Is there a general way to achieve the lower bound asymptotically?

A: Maximum likelihood estimation,

$$T_n^{\mathrm{MLE}}(X_1, \ldots, X_n) := \arg\min_{\theta \in \theta} \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(X_i).$$

The classical "Cramér-Rao theory" of variance:

(i) How small can variance be, among all unbiased estimators?

(ii) For which models can this lower bound be exactly achieved?

(iii) Is there a general way to achieve the lower bound asymptotically?

What if we are interested in quantities other than the variance?

An example of "stability" from veridical data science (Yu 2013, Yu-Kumbier 2020, Yu-Barter 2020), which"includes, but is much broader than, the concept of sampling variability"

II. Sensitivity

For an estimator $T_n : (\mathbb{R}^d)^n \to \mathbb{R}$ define its *sensitivity* via

$$\mathbb{S}_\theta(T_n) := \mathbb{E}_\theta \left[\sum_{i=1}^{n} \|\nabla_{x_i} T_n(X_1, \ldots, X_n)\|^2 \right].$$

Roughly speaking, this quantifies the total expected effect on T_n of slightly adjusting the samples $X_1, \ldots, X_n$.

Suppose $X_1, \ldots, X_n$ are latent variables of interest, and
the variables $X'_1, \ldots, X'_n$ are some measurements thereof.

Typically $X'_i := X_i + \varepsilon \cdot \xi_i$ for all $1 \leq i \leq n$, where $\xi_1, \ldots, \xi_n$ are i.i.d.
samples from $\mathcal{N}(0, I_d)$ which are also independent of $X_1, \ldots, X_n$.

In practice we compute $T_n(X'_1, \ldots, X'_n)$ and not $T_n(X_1, \ldots, X_n)$

How different are the resulting estimates?

Lots of other settings where data are perturbed by small noise
(Caroll-Stefanski 1990, Breiman 1996, Nagler 2018)

Note $\mathbb{V}_\theta(T_n)$ represents "stability under resampling":

$$\mathbb{V}_\theta(T_n) = \frac{1}{2}\mathbb{E}_\theta\left[|T_n(X_1,\ldots,X_n) - T_n(X'_1,\ldots,X'_n)|^2\right]$$

where $X'_1,\ldots,X'_n$ are i.i.d. samples, independent of $X_1,\ldots,X_n$.

Then, $\mathbb{S}_\theta(T_n)$ represents "stability under infinitesimal perturbation":

$$\mathbb{S}_\theta(T_n) \leftarrow \mathbb{E}_\theta\left[\left|\frac{T_n(X_1,\ldots,X_n) - T_n(X'_1,\ldots,X'_n)}{\varepsilon}\right|^2\right] \quad \text{as} \quad \varepsilon \to 0,$$

where $X'_i := X_i + \varepsilon \cdot \xi_i$ for all $1 \le i \le n$ as above.

Relation to robust statistics (Huber-Ronchetti 2009)? No!

- Robustness: proportion of samples are perturbed arbitrarily
- Sensitivity: all of the samples are perturbed slightly

Basic calculations for $X_1, \ldots, X_n \in \mathbb{R}$:

If T_n is sample mean, then $\nabla_{x_i} T_n(X_1, \ldots, X_n) = 1/n$ for all i, so

$$\mathbb{S}_\theta(T_n) = \mathbb{E}_\theta \left[\sum_{i=1}^n |\nabla_{x_i} T_n|^2 \right] = \mathbb{E}_\theta \left[\sum_{i=1}^n \left(\frac{1}{n} \right)^2 \right] = \frac{1}{n}.$$

If T_n is sample median, then $\nabla_{x_i} T_n(X_1, \ldots, X_n) = \mathbb{1}\{X_i = X_{((n+1)/2)}\}$

$$\mathbb{S}_\theta(T_n) = \mathbb{E}_\theta \left[\sum_{i=1}^n |\nabla_{x_i} T_n|^2 \right] = 1.$$

Example. Gaussian location family, $P_\theta = \mathcal{N}(\theta, 1)$

Note that $\mathbb{S}_\theta(T_n)$ is the Dirichlet energy of T_n under P_θ, so:

$$\mathbb{S}_\theta(T_n) \underset{\substack{\text{Gaussian}\\ \text{Poincaré}\\ \text{inequality}}}{\geq} \mathbb{V}_\theta(T_n) \underset{\substack{\text{Cramér}\\ \text{Rao}\\ \text{bound}}}{\geq} \frac{1}{n}.$$

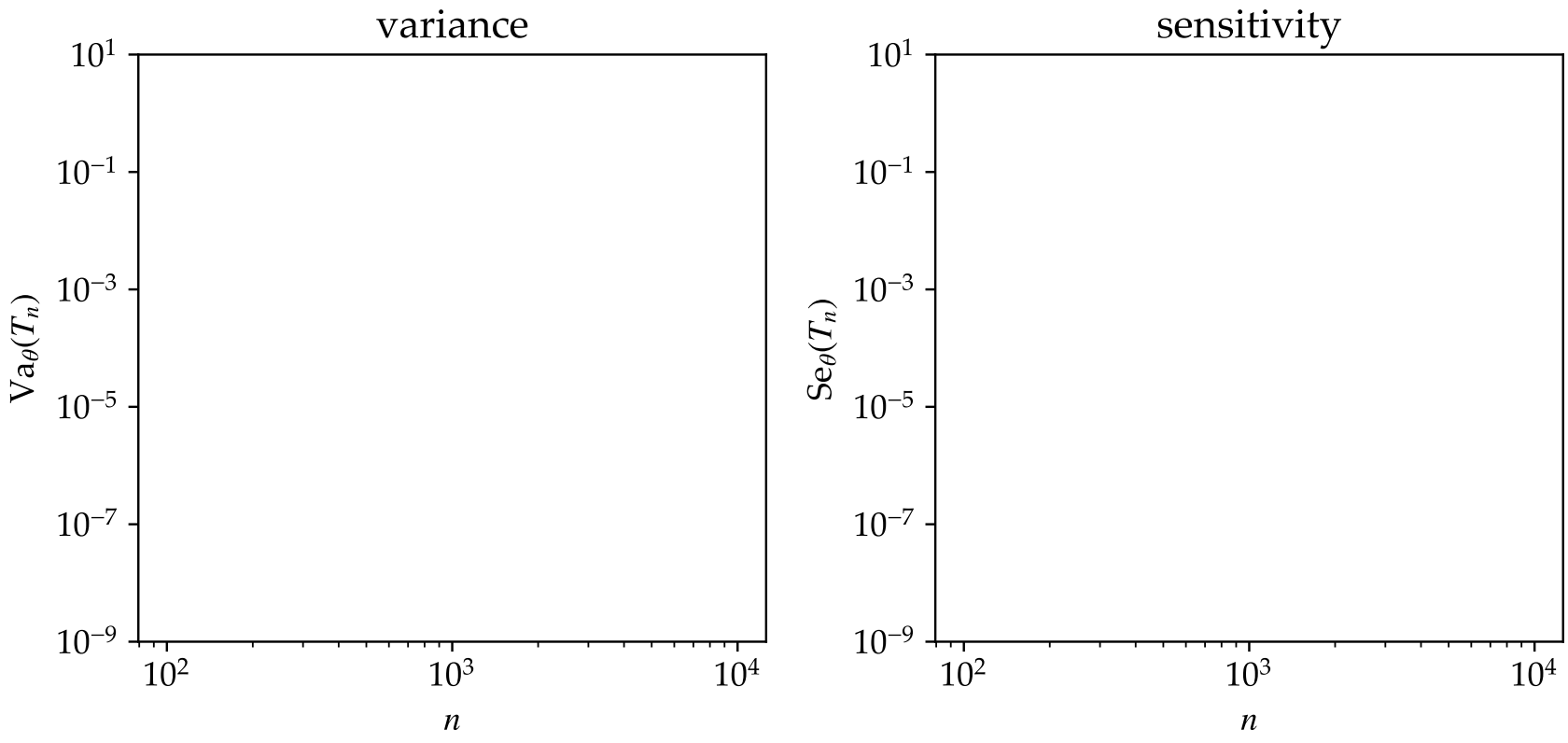
The sample mean has optimal variance and sensitivity!

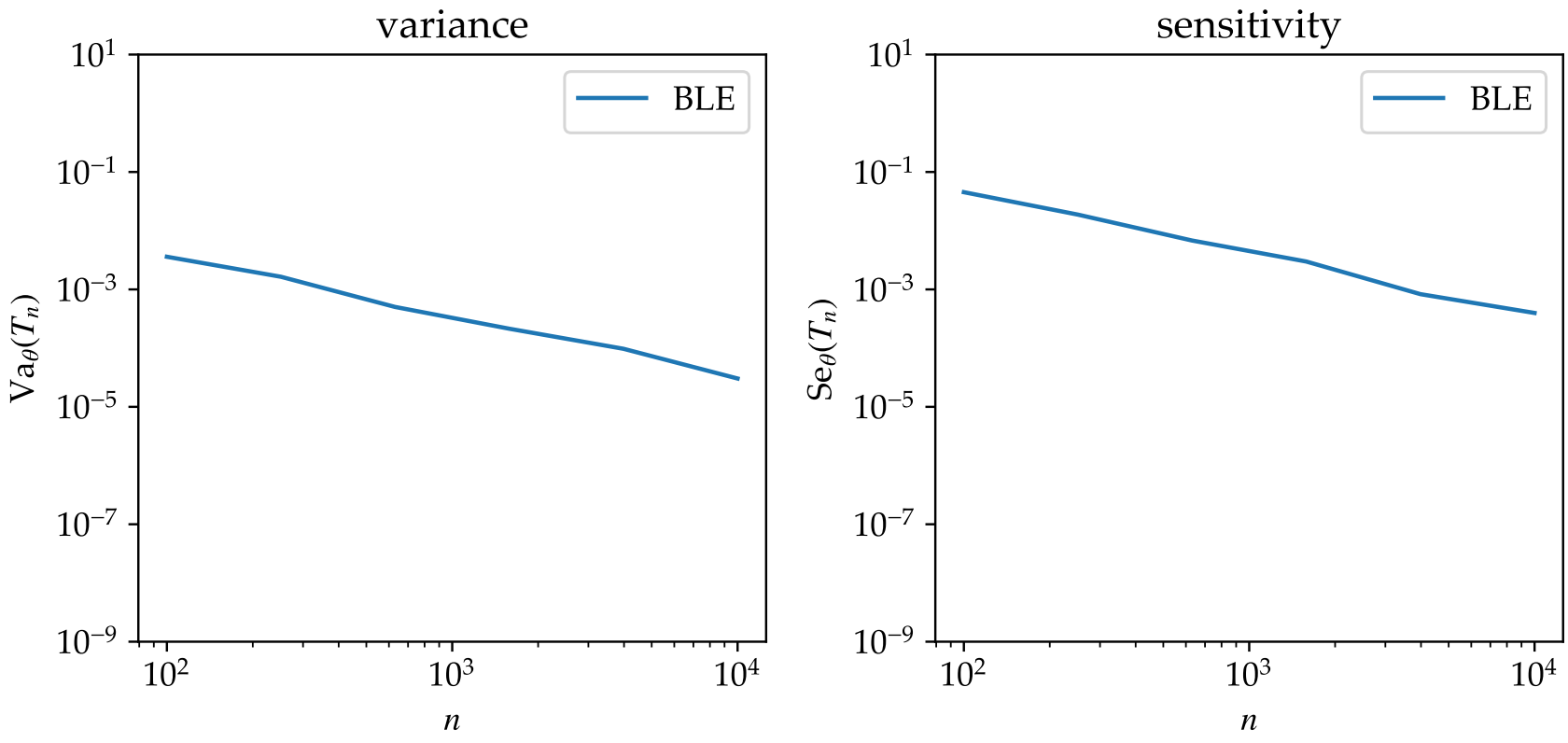
Bounds based on Poincaré inequality are not good enough in general...

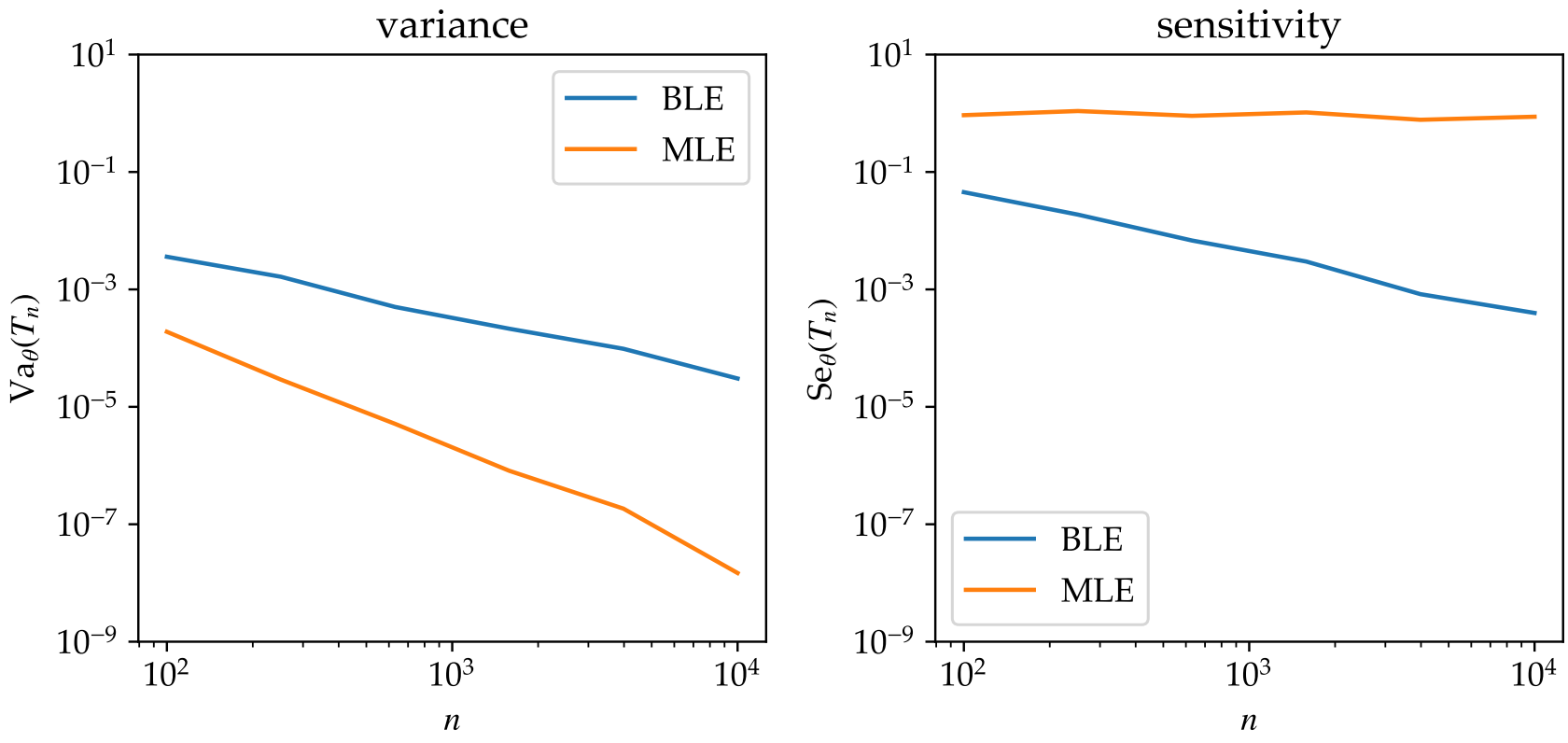
Example. Uniform scale family, $P_\theta = \mathrm{U}[0, \theta]$

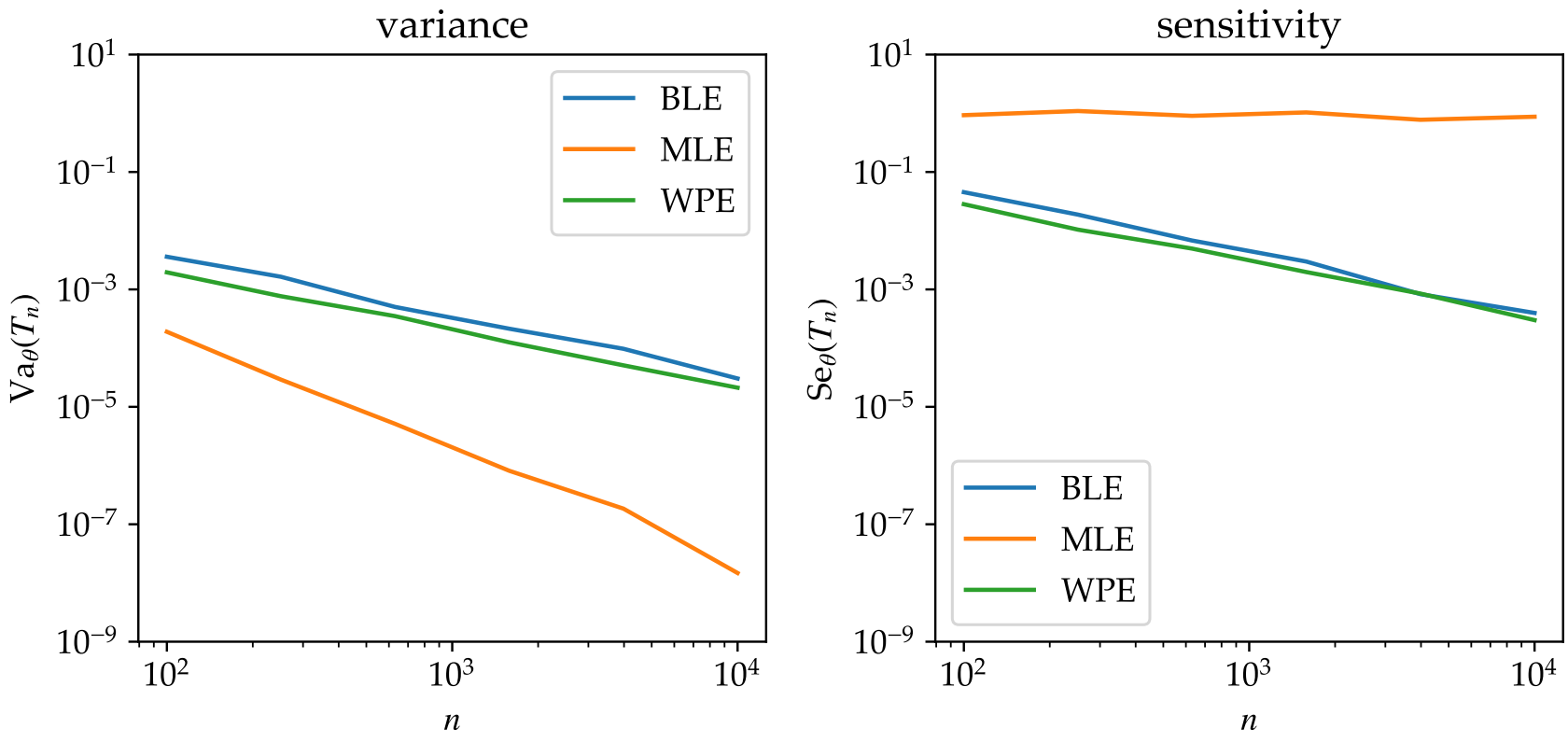
Natural estimators:

$$T_n^{\mathrm{MLE}} := \max_{1 \le i \le n} X_i \qquad T_n^{\mathrm{BLE}} := \frac{2}{n} \sum_{i=1}^{n} X_i$$


variance
sensitivity
10^{1}
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
$\mathrm{Va}_\theta(T_n)$
$\mathrm{Se}_\theta(T_n)$
10^{2}
10^{3}
10^{4}
n


variance
BLE
10^{1}
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
$\mathrm{Va}_{\theta}(T_n)$
10^{2}
10^{3}
10^{4}
n
sensitivity
BLE
10^{1}
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
$\mathrm{Se}_{\theta}(T_n)$
10^{2}
10^{3}
10^{4}
n


variance
sensitivity
BLE
MLE
10^{1}
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{2}
10^{3}
10^{4}
n
$\mathrm{Va}_\theta(T_n)$
$\mathrm{Se}_\theta(T_n)$


variance
sensitivity
BLE
MLE
WPE
10^{1}
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{2}
10^{3}
10^{4}
n
$\mathrm{Va}_\theta(T_n)$
$\mathrm{Se}_\theta(T_n)$

Example. Uniform scale family, $P_\theta = \mathrm{U}[0, \theta]$

Natural estimators:

$$T_n^{\mathrm{MLE}} := \max_{1 \le i \le n} X_i \qquad T_n^{\mathrm{BLE}} := \frac{2}{n} \sum_{i=1}^{n} X_i$$

Estimator with asymptotically optimal sensitivity:

$$T_n^{\mathrm{WPE}} := \frac{3}{2n^2} \sum_{i=1}^{n} (2i - 1) X_{(i)}$$

III. Wasserstein-Cramér-Rao Bound

Q: How small can sensitiity be, among all unbiased estimators?

A: Wasserstein-Cramér-Rao bound, stating

$$\mathbb{S}_\theta(T_n) \geq \frac{1}{nJ(\theta)}$$

for the Wasserstein information

$$J(\theta) := \mathbb{E}_\theta \left[\|\Phi_\theta(X)\|^2 \right].$$

Recall the proof of the usual Cramér-Rao bound

For sufficiently regular $\{p_\theta\}_{\theta\in\Theta\subseteq\mathbb{R}}$ and $T_1 : \mathbb{R}^d \to \mathbb{R}$, compute:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\theta[T_1] &= \frac{\mathrm{d}}{\mathrm{d}\theta}\int_{\mathbb{R}^d} T_1(x)p_\theta(x)\mathrm{d}\lambda(x) \\
&= \int_{\mathbb{R}^d} T_1(x)\frac{\mathrm{d}}{\mathrm{d}\theta}p_\theta(x)\mathrm{d}\lambda(x) \\
&= \int_{\mathbb{R}^d} T_1(x)\frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(x)\, p_\theta(x)\mathrm{d}\lambda(x) \\
&= \int_{\mathbb{R}^d} (T_1(x)-\theta)\frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(x)\, p_\theta(x)\mathrm{d}\lambda(x) \\
&\leq \sqrt{\int_{\mathbb{R}^d}(T_1(x)-\theta)^2 p_\theta(x)\mathrm{d}\lambda(x)\int_{\mathbb{R}^d}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(x)\right)^2\mathrm{d}\lambda(x)},
\end{aligned}$$

then rearrange.

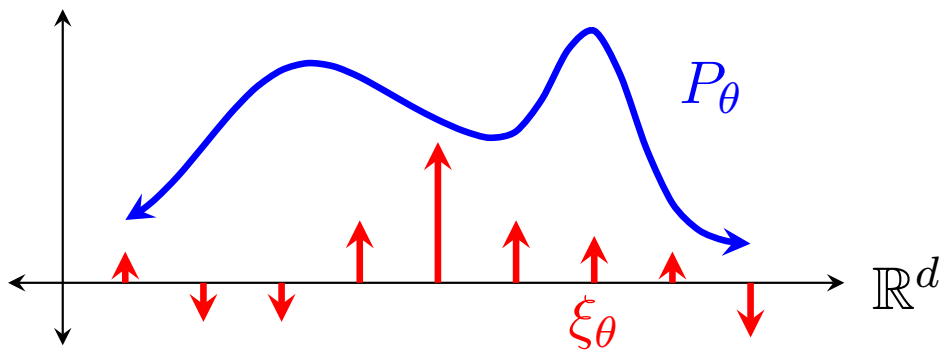
Differentiating under the integral is weak form of the *reaction equation*

$$\partial_\theta P_\theta = \xi_\theta P_\theta$$

applied to the test function T_1, namely

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \int_{\mathbb{R}^d} T_1(x) \mathrm{d}P_\theta(x) = \int_{\mathbb{R}^d} T_1(x) \xi_\theta(x) \mathrm{d}P_\theta(x)$$

Geometrically:



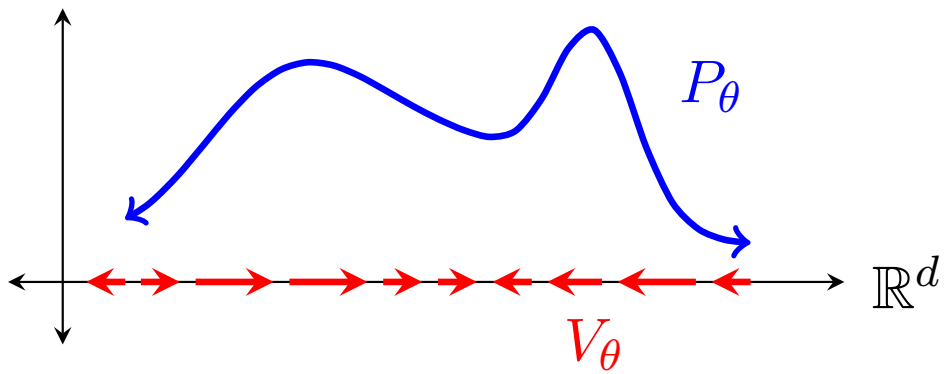
Instead, we could have applied the *continuity equation*:

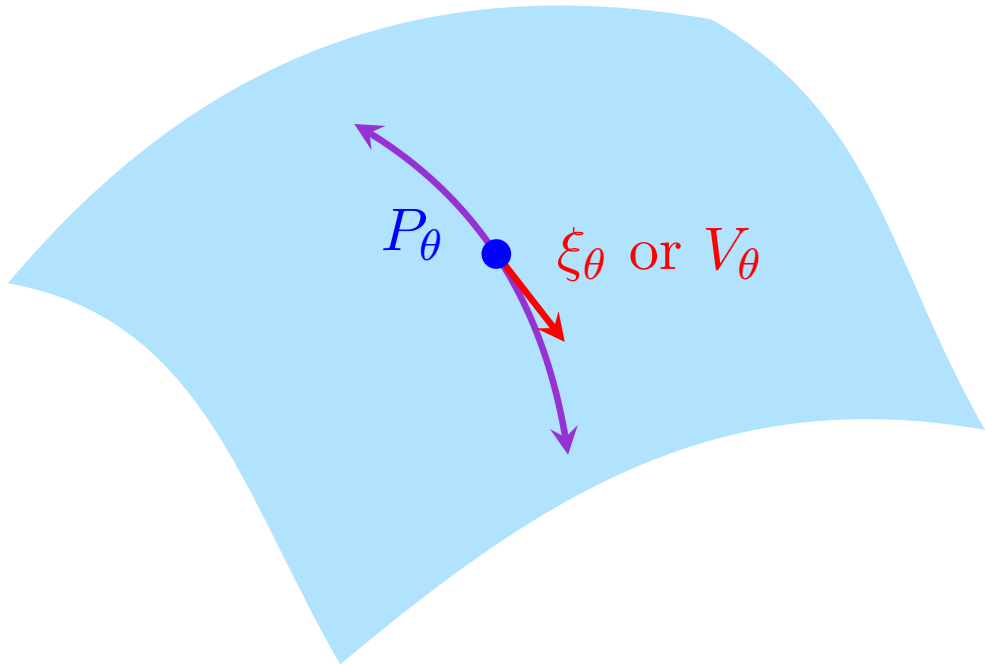
$$\partial_\theta P_\theta + \operatorname{div}(V_\theta P_\theta) = 0$$

to the test function T_1, namely

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \int_{\mathbb{R}^d} T_1(x) \mathrm{d}P_\theta(x) = \int_{\mathbb{R}^d} (\nabla T_1(x))^\top V_\theta(x) \mathrm{d}P_\theta(x)$$

Geometrically:




P_θ
ξ_θ or V_θ

Need to guarantee that $\mathcal{P}$ is regular enough to satisfy the continuity equation along all one-dimensional submodels:

Definition (García Trillos-AQJ-Sen 2025)

A model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is called *differentiable in the Wasserstein sense (DWS)* at $\theta \in \Theta$ if there exists a function $\Phi_\theta : \mathbb{R}^d \to \mathbb{R}^{d\times p}$ with

$$\int_{\mathbb{R}^d} \left\| \mathbf{t}_{P_\theta \to P_{\theta+th}}(x) - x - t\Phi_\theta(x)h \right\|^2 \mathrm{d}P_\theta(x) = o(t^2)$$

as $t \to 0$. We say Φ_θ is its *transport linearization*, and we define its *Wasserstein information matrix* as

$$J(\theta) = \mathbb{E}_\theta \left[(\Phi_\theta(X))^\top \Phi_\theta(X) \right].$$

Example. Location family, $P_\theta = P_0(\,\cdot - \theta)$ for fixed P_0, and $\theta \in \mathbb{R}^d$

Note $\mathbf{t}_{P_{\theta_0} \to P_{\theta_1}}(x) = x + \theta_1 - \theta_0$. Thus, $\Phi_\theta \equiv I_d$ hence $J(\theta) = I_d$.

Example. Scale family, $P_\theta = P_1(\,\cdot\,/\,\theta)$ for fixed P_0, and $\theta > 0$

Note $\mathbf{t}_{P_{\theta_0} \to P_{\theta_1}}(x) = (\theta_1/\theta_0)x$, so $\Phi_\theta(x) = x/\theta$ and $J(\theta) = \int \|x\|^2 \mathrm{d}P_1(x)$.

Example. Gaussian correlation family, $P_\theta = \mathcal{N}(0, \Sigma_\theta)$, where

$$\Sigma_\theta := \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix}$$

for $-1 < \theta < 1$. Some direct calculations (Olkin-Pukelsheim 1982) give

$$\Phi_\theta(x) = \frac{1}{2(1-\theta^2)} \begin{pmatrix} -\theta & 1 \\ 1 & -\theta \end{pmatrix} x$$

$J(\theta) = 1 - \theta^2$.

Analogous to classical case:

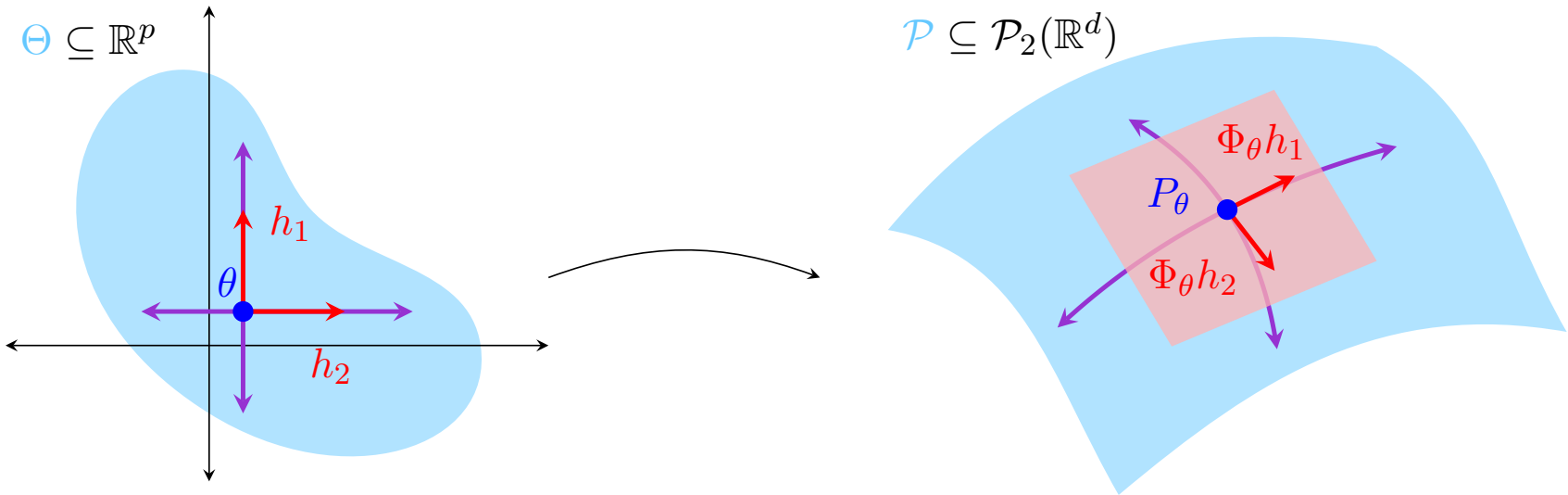
$$\text{DWS:} \qquad \int_{\mathbb{R}^d} \left\| \mathbf{t}_{P_\theta \to P_{\theta+th}}(x) - x - t\Phi_\theta(x)h \right\|^2 \mathrm{d}P_\theta(x) = o(t^2)$$

$$\text{DQM:} \qquad \int_{\mathbb{R}^d} \left| \sqrt{\frac{\mathrm{d}P_{\theta+th}}{\mathrm{d}P_\theta}(x)} - 1 - t(G_\theta(x))^\top h \right|^2 \mathrm{d}P_\theta(x) = o(t^2)$$

Riemannian interpretation, as in classical case (Amari 1985):

$$\int_{\mathbb{R}^d} \left\| \mathrm{Log}_{P_\theta}(P_{\theta+th}) - t\Phi_\theta(x)h \right\|^2 \mathrm{d}P_\theta(x) = o(t^2),$$

meaning Φ_θ is the differential of $P : \Theta \to \mathcal{P}$ at θ.


$\Theta \subseteq \mathbb{R}^p$
h_1
θ
h_2
$\mathcal{P} \subseteq \mathcal{P}_2(\mathbb{R}^d)$
$\Phi_\theta h_1$
P_θ
$\Phi_\theta h_2$

Theorem (García Trillos-AQJ-Sen 2025)

Suppose that $\mathcal{P} = \{P_\theta : \theta \in \Theta\} \subseteq \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ *is DWS,* $\chi : \Theta \to \mathbb{R}^k$ *is differentiable and* $T_n : (\mathbb{R}^d)^n \to \mathbb{R}^k$ *is an unbiased estimator of* $\chi(\theta)$ *for all* $\theta \in \Theta$*. Under some technical assumptions, we have*

$$\mathbb{S}_\theta(T_n) \succeq \frac{1}{n}(D\chi(\theta))^\top (J(\theta))^{-1} D\chi(\theta),$$

for all $\theta \in \Theta$

Extends some earlier work in differential geometry (Li-Zhao 2023)

Example. Location family, $\mathbb{S}_\theta(T_n) \succeq I_d/n$

Example. Scale family, $\mathbb{S}_\theta(T_n) \geq 1/(n \int \|x\|^2 \mathrm{d}P_1(x))$

IV. Achieving the Lower Bound

Q: For which models can this lower bound be exactly achieved?

A: Transport families,

$$\Phi_\theta(x) = D\phi(x)(\Lambda(\theta))^{-1}(D\chi(\theta))^\top$$

where

$$\Lambda(\theta) := \mathbb{E}_\theta[(D\phi(X))^\top D\phi(X)].$$

Definition (García Trillos-AQJ-Sen 2025)

Suppose that $\mathcal{P}$ is DWS, that $J(\theta)$ is invertible for each $\theta \in \Theta$, and that $T_n : (\mathbb{R}^d)^n \to \mathbb{R}^k$ is an unbiased estimator of $\chi(\theta)$, for some differentiable function $\chi : \Theta \to \mathbb{R}^k$. We say that T_n is *sensitivity-efficient* if we have

$$\mathbb{S}_\theta(T_n) = \frac{1}{n}(D\chi(\theta))^\top (J(\theta))^{-1} D\chi(\theta), \tag{1}$$

for all $\theta \in \Theta$.

Definition (García Trillos-AQJ-Sen 2025)

For a locally Lipschitz function $\phi : \mathbb{R}^d \to \mathbb{R}^k$ and a differentiable function $\chi : \Theta \to \mathbb{R}^k$, a *transport family* is a DWS model $\mathcal{P} = \{P_\theta : \theta \in \Theta\} \subseteq \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ for open $\Theta \subseteq \mathbb{R}^p$ if we have

$$\Phi_\theta(x) = D\phi(x)(\Lambda(\theta))^{-1}(D\chi(\theta))^\top$$

for Lebesgue almost every $x \in \mathbb{R}^d$, for all $\theta \in \Theta$, and where $\Lambda(\theta) := \mathbb{E}_\theta[(D\phi(X))^\top D\phi(X)]$ is assumed to be invertible. The function ϕ is called the *potential*, and the function χ is called the *parameterization*.

Theorem (García Trillos-AQJ-Sen 2025)

Under some technical assumptions, $\mathcal{P}$ is a transport familty with parameterization χ if and only if there exists an unbiased sensivity-efficient estimator of $\chi(\theta)$,in which case the estimator may be taken to be

$$T_n(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \phi(X_i)$$

where ϕ is the potential of $\mathcal{P}$.

Proof via equality in Cauchy-Schwarz, as classical case (Wijsman 1973)

Allows us to handle many examples...

Example. Location family, $P_\theta = P_0(\,\cdot - \theta)$ for fixed P_0, and $\theta \in \mathbb{R}^d$

We have $\Phi_\theta \equiv I_d$, thus $D\phi(x) = I_d$ hence $\phi(x) = x$. So, sample mean $T_n = \sum_{i=1}^n X_i/n$ is unbiased and sensitivity-efficient estimator of θ.

Example. Scale family, $P_\theta = P_1(\,\cdot\,/\,\theta)$ for fixed P_0, and $\theta > 0$

We have $\Phi_\theta(x) = x/\theta$, thus $\nabla\phi(x) = x$ hence $\phi(x) = \|x\|^2/2$. So, sample second moment $T_n = \sum_{i=1}^n \|X_i\|^2/n$ is unbiased and sensitivity-efficient estimator of θ^2.

Example. Linear regression, $X = \mathbf{W}\theta + \varepsilon$ for $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_d)$.

Some further calculations show this is a transport family, and the ordinary least squares (OLS) estimator $T_n = (\mathbf{W}^\top \mathbf{W})^{-1}\mathbf{W}^\top X$ is an unbiased and sensitivity-efficient estimator of θ.

Q: Is there a general way to achieve the lower bound asymptotically?

A: Wasserstein projection estimation (WPE),

$$T_n^{\mathrm{WPE}}(X_1, \ldots, X_n) := \arg\min_{\theta \in \theta} W_2^2 \left(P_\theta, \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} \right)$$

Definition (García Trillos-AQJ-Sen 2025)

Suppose that $\mathcal{P}$ is DWS, and that $T_n : (\mathbb{R}^d)^n \to \mathbb{R}^k$ is an estimator of $\chi(\theta)$. We say that T_n is *asymptotically sensitivity-efficient* if we have

$$n \sum_{i=1}^{n} (D_{x_i} T_n(X_1, \ldots, X_n))^\top D_{x_i} T_n(X_1, \ldots, X_n) \\ \to (D\chi(\theta))^\top (J(\theta))^{-1} D\chi(\theta)$$

in probability, when $X_1, X_2, \ldots$ are i.i.d. from P_θ, for each $\theta \in \Theta$.

Note convergence in probability not convergence in expectation

Define $\mathcal{L}(x_1, \ldots, x_n, \theta) := W_2^2(P_\theta, \bar{P}_n)$ and note that differentiating the first-order conditions for $\hat{\theta}_n := T_n^{\mathrm{WPE}}$ gives (formally):

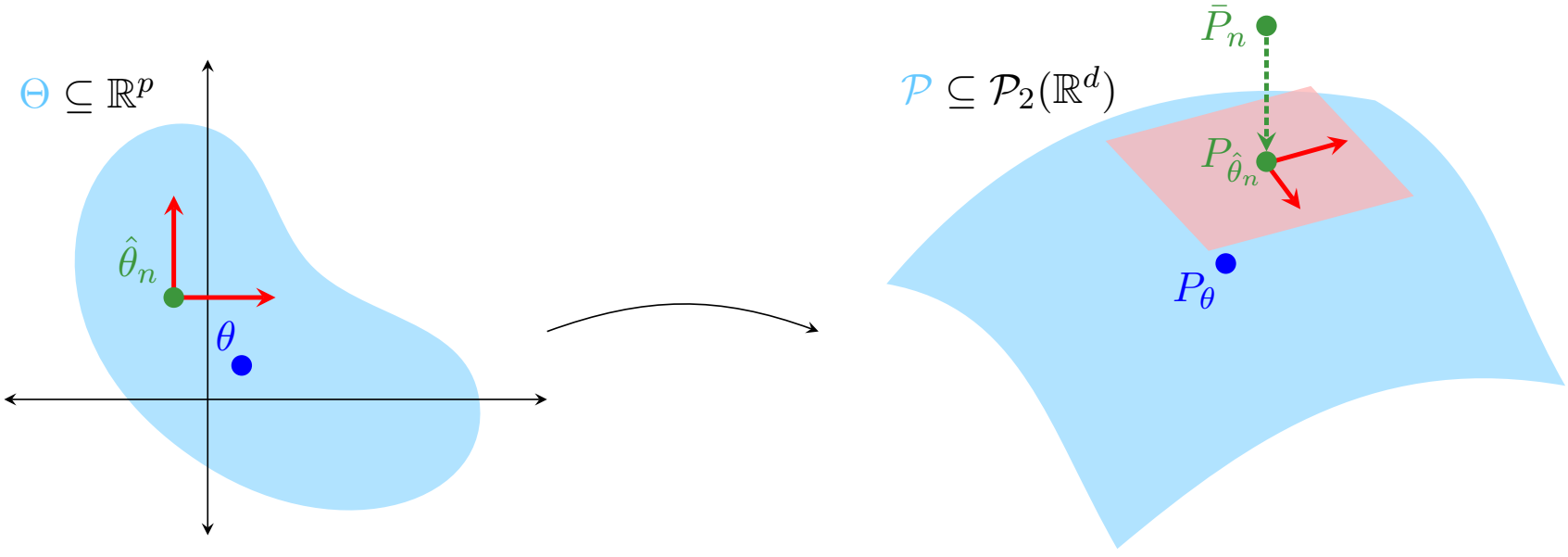
$$0 = \nabla_{x_i} \frac{\partial}{\partial \theta} \mathcal{L}(X_1, \ldots, X_n, \hat{\theta}_n) + \frac{\partial^2}{\partial \theta^2} \mathcal{L}(X_1, \ldots, X_n, \hat{\theta}_n) \nabla_{x_i} \hat{\theta}_n.$$

Rearranging and summing gives:

$$\begin{aligned} & n \sum_{i=1}^{n} \left\| \nabla_{x_i} \hat{\theta}_n(X_1, \ldots, X_n) \right\|^2 \\ & = \left(\frac{\partial^2}{\partial \theta^2} \mathcal{L}(X_1, \ldots, X_n, \hat{\theta}_n) \right)^{-2} \frac{1}{n} \sum_{i=1}^{n} \left\| n \nabla_{x_i} \frac{\partial}{\partial \theta} \mathcal{L}(X_1, \ldots, X_n, \hat{\theta}_n) \right\|^2 . \end{aligned}$$

Just need to show that $J(\theta)$ is the limit of both

$$\frac{\partial^2}{\partial \theta^2} \mathcal{L}(X_1, \ldots, X_n, \hat{\theta}_n) \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} \left\| n \nabla_{x_i} \frac{\partial}{\partial \theta} \mathcal{L}(X_1, \ldots, X_n, \hat{\theta}_n) \right\|^2 .$$


$\Theta \subseteq \mathbb{R}^p$
$\hat{\theta}_n$
θ
$\mathcal{P} \subseteq \mathcal{P}_2(\mathbb{R}^d)$
$\bar{P}_n$
$P_{\hat{\theta}_n}$
P_θ

When $X_1, \ldots, X_n$ live in dimension $d = 1$, we can make this argument rigorous using representation of transport maps as quantile functions:

Theorem (García Trillos-AQJ-Sen 2025)

Suppose $\mathcal{P} = \{P_\theta : \theta \in \Theta\} \subseteq \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R})$ is DWS, and that some mild technical assumptions hold. If $X_1, X_2, \ldots$ are i.i.d. samples from P_{θ^*} and T_n^{WPE} is a consistent estimator of $\theta^* \in \Theta$, then we have

$$n \sum_{i=1}^{n} \left(\frac{\partial T_n^{\mathrm{WPE}}}{\partial x_i}(X_1, \ldots, X_n) \right) \left(\frac{\partial T_n^{\mathrm{WPE}}}{\partial x_i}(X_1, \ldots, X_n) \right)^{\top} \to (J(\theta^*))^{-1}$$

in probability as $n \to \infty$, for each $\theta^* \in \Theta$.

Example. Uniform scale family, $P_\theta = \mathrm{U}[0, \theta]$

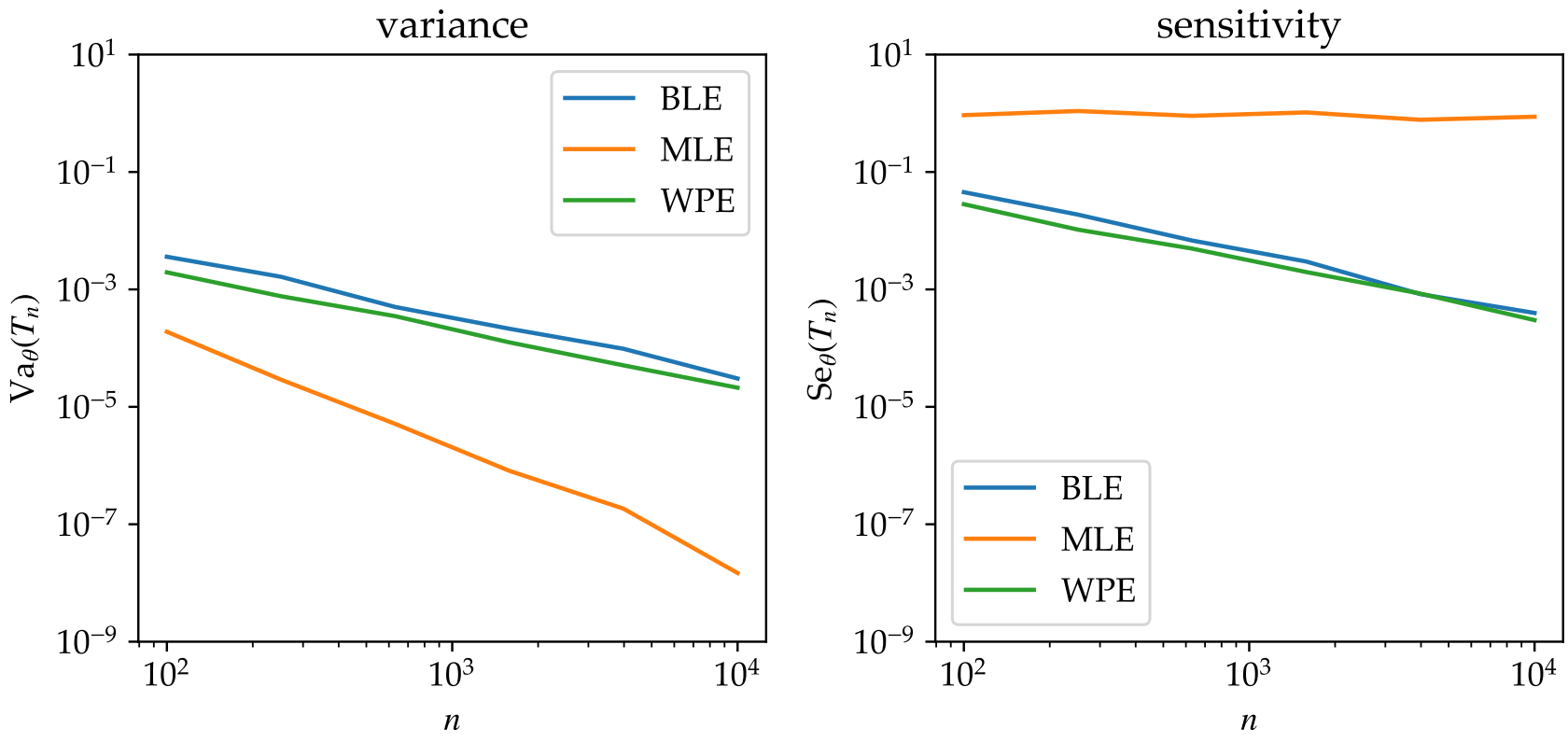
Recall that $\sum_{i=1}^n X_i^2/n$ is a sensitivity-efficient estimator of θ^2, and there is no sensitivity-efficient estimator of θ. Asymptotically?

Directly compute:

$$W_2^2(P_\theta, \bar{P}_n) = \int_0^1 \left|\bar{F}_n^{-1}(u) - \theta u\right|^2 \mathrm{d}u,$$

hence

$$T_n^{\mathrm{WPE}}(X_1, \ldots, X_n) = \frac{\int_0^1 u\, \bar{F}_n^{-1}(u)\mathrm{d}u}{\int_0^1 u^2 \mathrm{d}u} = \frac{3}{2n^2} \sum_{i=1}^n (2i-1) X_{(i)}.$$


variance
sensitivity
BLE
MLE
WPE
10^{1}
10^{-1}
10^{-3}
10^{-5}
10^{-7}
10^{-9}
10^{2}
10^{3}
10^{4}
n
$\mathrm{Va}_\theta(T_n)$
$\mathrm{Se}_\theta(T_n)$

When $X_1, \ldots, X_n$ live in dimension $d \geq 2$, we need stronger assumptions in order to make this argument rigorous.

Theorem (García Trillos-AQJ-Sen 2025)

Suppose $\mathcal{P} = \{P_\theta : \theta \in \Theta\} \subseteq \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ is DWS, and that some strong technical assumptions hold. If $X_1, X_2, \ldots$ are i.i.d. samples from P_{θ^*} and T_n^{WPE} is a consistent estimator of $\theta^* \in \Theta$, then we have

$$n \sum_{i=1}^{n} \left(\nabla_{x_i} T_n^{\mathrm{WPE}}(X_1, \ldots, X_n)\right)^\top \left(\nabla_{x_i} T_n^{\mathrm{WPE}}(X_1, \ldots, X_n)\right) \to (J(\theta^*))^{-1}$$

in probability as $n \to \infty$, for each $\theta^* \in \Theta$.

V. Outlook

The "Wasserstein-Cramér-Rao theory" of sensitivity:

(i) How small can sensitivity be, among all unbiased estimators?

(ii) For which models can this lower bound be exactly achieved?

(iii) Is there a general way to achieve the lower bound asymptotically?

Ongoing work:

- How to define sensitivity in discrete statistical models?
- How to make sense of semiparametric sensitivity-efficiency?
- How to tradeoff between optimizing variance and sensitivity?

What other measures of instability can be studied by suitable geometries on spaces of probability meaures?

Thank you!

References

S.-I. Amari. *Differential-Geometrical Methods in Statistics*. Lecture Notes in Statistics, 28. Springer-Verlag, New York, 1985.

L. Breiman. Heuristics of instability and stabilization in model selection. *Ann. Statist.*, 24:2350-2383, 1996.

R. J. Carroll and L. A. Stefanski. Approximate quasi-likelihoo eEstimation in models with surrogate predictors. *J. Amer. Stat. Assoc.*, 85:652-663, 1990

P. J. Huber and E. M. Ronchetti. *Robust statistics*, 2nd ed. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, 2009

W. Li, W. an J. Zhao. Wasserstein information matrix. *Inf. Geom.*, 6:203–255, 2023.

T. Nagler. Asymptotic Analysis of the jittering kernel density estimator. *Math. Methods Stat.*, 27:32-46, 2018.

I. Olkin and F. Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra Appl.*, 48:257-263, 1982.

S. Rachev and L. Rüschendorf. *Mass Transportation Problems*. Volume 1 of *Probability and its Applications*. Springer-Verlag, New York, 1998.

A. W. van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, 3. Cambridge University Press, Cambridge, UK, 1998.

C. Villani. *Optimal Transport: Old and New*. Volume 338 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin, 2009.

R. A. Wijsman. On the Attainment of the Cramer-Rao lower bound. *Ann. Statist.*, 1:538-542, 1973.

B. Yu. Stability. *Bernoulli*, 19:1484-1500, 2013.

B. Yu and R. Barter.The datam science process: one culture. *J. Amer. Stat. Assoc.*, 115:672–674, 2020

B. Yu and K. Kumbier. Veridical data science. *Proc. Natl. Acad. Sci. U.S.A.*, 117:3920-3929, 2020.