A Kernel Test of Goodness of Fit

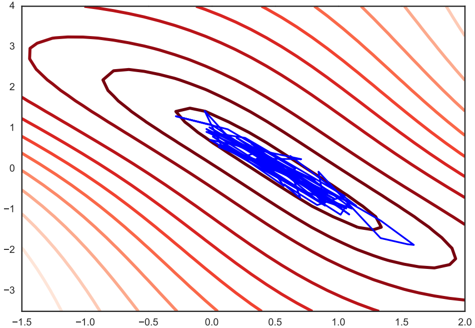
Kacper Chwialkowski, Heiko Strathmann, Arthur Gretton

June 21, 2016

Gatsby Unit and CS Department, UCL

Motivating example: testing output of approximate MCMC

Approximate MCMC: tradeoff between bias and computation
(e.g. *Austerity in MCMC Land* [2])

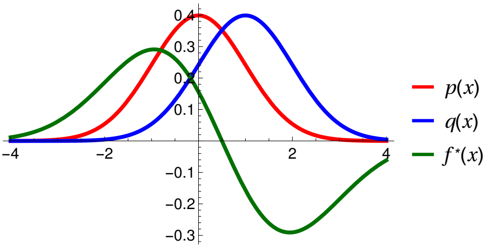


$\theta_1 \sim \mathcal{N}(0, 10); \theta_2 \sim \mathcal{N}(0, 1)$
$X_i \sim \frac{1}{2}\mathcal{N}(\theta_1, 4) + \frac{1}{2}\mathcal{N}(\theta_1 + \theta_2, 4).$

How to check if MCMC samples match target distribution?

Maximum mean discrepancy: metric between p and q

$$MMD(p, q, F) = \sup_{\|f\|_F < 1} [\mathbb{E}_q f - \mathbb{E}_p f]$$

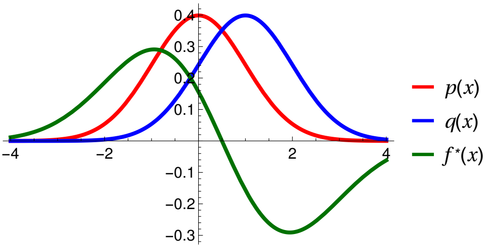


- F is an Reproducing Kernel Hilbert Space.
- f^* is the function that attains the supremum.

Can we compute MMD when q are MCMC samples, p is model?

Maximum mean discrepancy: metric between p and q

$$MMD(p, q, F) = \sup_{\|f\|_F < 1} [\mathbb{E}_q f - \mathbb{E}_p f]$$



- F is an Reproducing Kernel Hilbert Space.
- f^* is the function that attains the supremum.

Can we compute MMD when q are MCMC samples, p is model?

Problem: don't have $\mathbb{E}_p f$ in closed form

Main idea (by Stein)

To get rid of $\mathbb{E}_p f$ in

$$\sup_{\|f\|_F < 1} [\mathbb{E}_q f - \mathbb{E}_p f]$$

we will use the cornerstone of modern ML

Main idea (by Stein)

To get rid of $\mathbb{E}_p f$ in

$$\sup_{\|f\|_F < 1} [\mathbb{E}_q f - \mathbb{E}_p f]$$

we will use the cornerstone of modern ML

Integration by parts

Main idea (by Stein)

To get rid of $\mathbb{E}_p f$ in

$$\sup_{\|f\|_F<1} [\mathbb{E}_q f - \mathbb{E}_p f]$$

we will use the cornerstone of modern ML

Integration by parts

Define the **Stein operator**

$$T_p f = f' + \log' p \cdot f$$

Then

$$\mathbb{E}_p T_p f = 0$$

Maximum Stein Discrepancy

Stein operator

$$T_p f = f' + \log' p \cdot f$$

Maximum Stein Discrepancy (MSD)

$$MSD(p, q, F) = \sup_{\|g\|_{\mathcal{F}} < 1} \mathbb{E}_q T_p g - \mathbb{E}_p T_p g$$

Stein operator

$$T_p f = f' + \log' p \cdot f$$

Maximum Stein Discrepancy (MSD)

$$MSD(p, q, F) = \sup_{\|g\|_{\mathcal{F}} < 1} \mathbb{E}_q T_p g - \cancel{\mathbb{E}_p T_p g}$$

Stein operator

$$T_p f = f' + \log' p \cdot f$$

Maximum Stein Discrepancy (MSD)

$$MSD(p, q, F) = \sup_{\|g\|_{\mathcal{F}} < 1} \mathbb{E}_q T_p g - \cancel{\mathbb{E}_p T_p g} = \sup_{\|g\|_{\mathcal{F}} < 1} \mathbb{E}_q T_p g$$

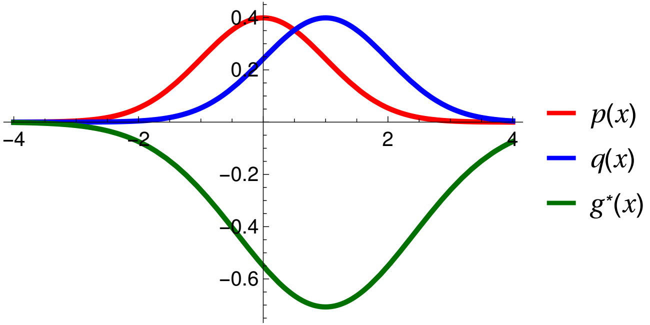
Maximum Stein Discrepancy

Stein operator

$$T_p f = f' + \log' p \cdot f$$

Maximum Stein Discrepancy (MSD)

$$MSD(p, q, F) = \sup_{\|g\|_{\mathcal{F}}<1} \mathbb{E}_q T_p g - \cancel{\mathbb{E}_p T_p g} = \sup_{\|g\|_{\mathcal{F}}<1} \mathbb{E}_q T_p g$$



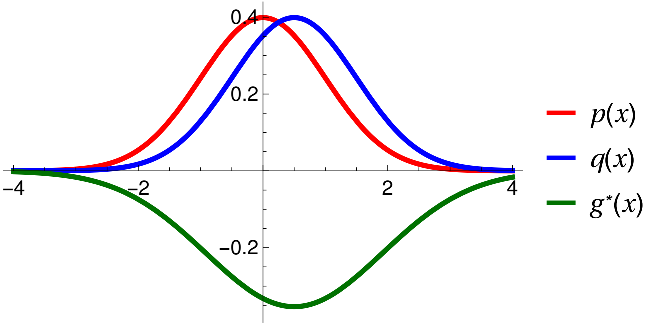
Maximum Stein Discrepancy

Stein operator

$$T_p f = f' + \log' p \cdot f$$

Maximum Stein Discrepancy (MSD)

$$MSD(p, q, F) = \sup_{\|g\|_{\mathcal{F}} < 1} \mathbb{E}_q T_p g - \cancel{\mathbb{E}_p T_p g} = \sup_{\|g\|_{\mathcal{F}} < 1} \mathbb{E}_q T_p g$$



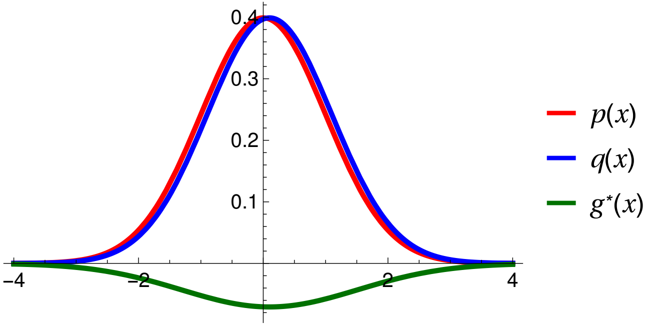
Maximum Stein Discrepancy

Stein operator

$$T_p f = f' + \log' p \cdot f$$

Maximum Stein Discrepancy (MSD)

$$MSD(p, q, F) = \sup_{\|g\|_{\mathcal{F}} < 1} \mathbb{E}_q T_p g - \cancel{\mathbb{E}_p T_p g} = \sup_{\|g\|_{\mathcal{F}} < 1} \mathbb{E}_q T_p g$$



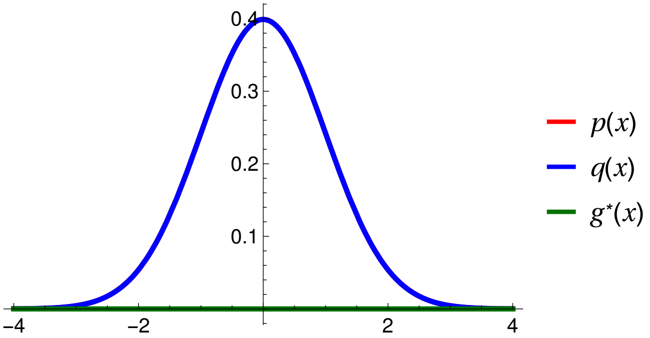
Maximum Stein Discrepancy

Stein operator

$$T_p f = f' + \log' p \cdot f$$

Maximum Stein Discrepancy (MSD)

$$MSD(p, q, F) = \sup_{\|g\|_{\mathcal{F}}<1} \mathbb{E}_q T_p g - \cancel{\mathbb{E}_p T_p g} = \sup_{\|g\|_{\mathcal{F}}<1} \mathbb{E}_q T_p g$$



Maximum Stein Discrepancy has simple closed-form expression

Closed-form expression for MSD: given $Z, Z' \sim q$, then

$$MSD(p, q, G) = \mathbb{E}_q h_p(Z, Z')$$

where

$$\begin{aligned} h_p(x, y) := {} & \partial_x \log p(x) \partial_x \log p(y) k(x, y) \\ & + \partial_y \log p(y) \partial_x k(x, y) \\ & + \partial_x \log p(x) \partial_y k(x, y) \\ & + \partial_x \partial_y k(x, y). \end{aligned}$$

and k is RKHS kernel for F

Only depends on kernel and $\partial_x \log p(x)$*.*
Do not need to normalize p*, or sample from it.*

Maximum Stein Discrepancy zero $\iff p = q$

Theorem

If the kernel k is C_0-universal, $\mathbb{E}_q h_q(Z,Z) < \infty$ and $\mathbb{E}_q \left(\log' \frac{p(Z)}{q(Z)}\right)^2 < \infty$ then

$$MSD(p, q, G) = 0 \text{ if and only if } p = q.$$

Kernel is C_0-universal if $f \to \int_X f(x)k(x,\cdot)d\mu(x)$ if is injective for all probability measures μ and all $f \in L^p(X,\mu)$, where $p \in [1,\infty]$.

The assumption $\mathbb{E}_q \left(\log' \frac{p(Z)}{q(Z)}\right)^2 < \infty$ states that difference between scores $\log' p$ and $\log' q$ is square integrable.

Empirical estimate of MSD: V-statistic

Empirical estimate of $\mathbb{E}_q h_p(Z, Z')$ is a *V-statistic*:

$$V_n(h_p) = \frac{1}{n^2} \sum_{i,j=1}^{n} h_p(Z_i, Z_j),$$

$\{Z_1, \ldots Z_t \ldots Z_n\}$ time series with marginal distrib. q

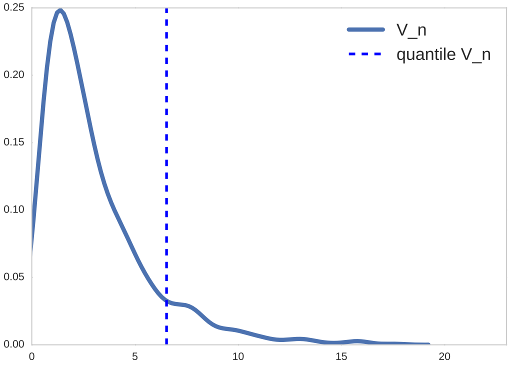
Empirical estimate of MSD: V-statistic

Empirical estimate of $\mathbb{E}_q h_p(Z, Z')$ is a *V-statistic*:

$$V_n(h_p) = \frac{1}{n^2} \sum_{i,j=1}^{n} h_p(Z_i, Z_j),$$

$\{Z_1, \ldots Z_t \ldots Z_n\}$ time series with marginal distrib. q

What are "typical" values of $\mathbb{E}_q h_p(Z, Z')$ *when* $p = q$ *?*



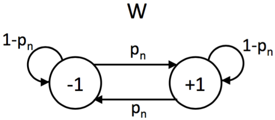
Distribution of statistic under null ($p = q$)

To estimate quantiles of $V_n(h_p)$ under the null (when $p = q$), we use **wild bootstrap**

$$B_n(h_p) = \frac{1}{n^2} \sum_{i,j=1}^{n} W_i W_j h_p(X_i, X_j).$$

where W_i are correlated zero mean RVs.

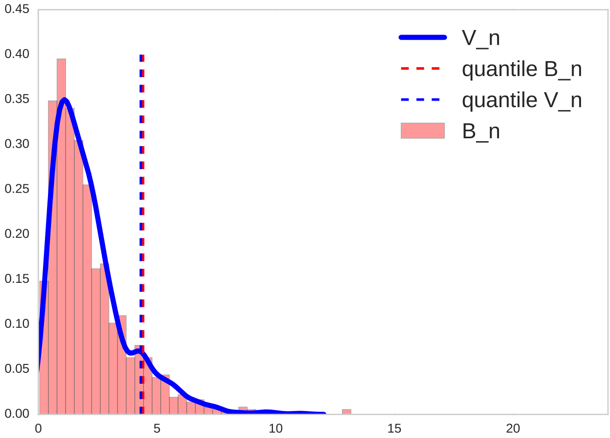
$$Cov(W_i, W_j) = (1 - 2p_n)^{-|i-j|}$$



p_n is the probability of the change and should be set to $o(n)$.

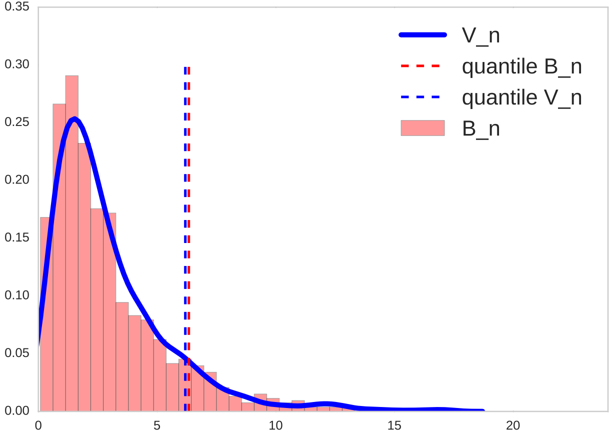
Wild bootstrapping; small correlation

$$X_t = 0.1X_{t-1} + \sqrt{1-0.1^2}\epsilon_t, \quad \epsilon_t \sim N(0,1)$$



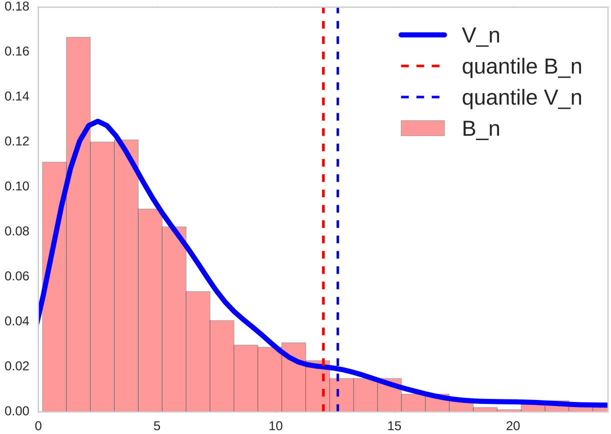
Wild bootstrapping, medium correlation

$$X_t = 0.4X_{t-1} + \sqrt{1 - 0.4^2}\epsilon_t, \quad \epsilon_t \sim N(0, 1)$$

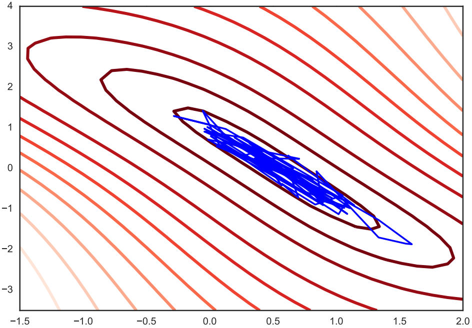


Wild bootstrapping; huge correlation

$$X_t = 0.7X_{t-1} + \sqrt{1-0.7^2}\epsilon_t, \quad \epsilon_t \sim N(0,1)$$

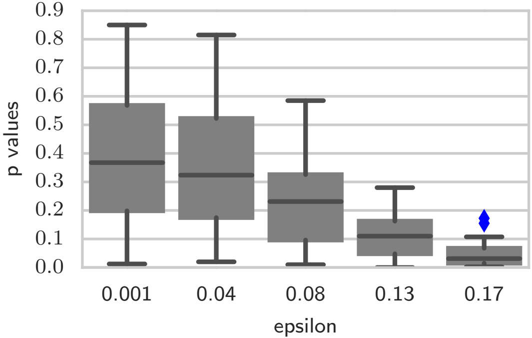


Approximate MCMC: tradeoff between bias and computation
(e.g. *Austerity in MCMC Land* [2])



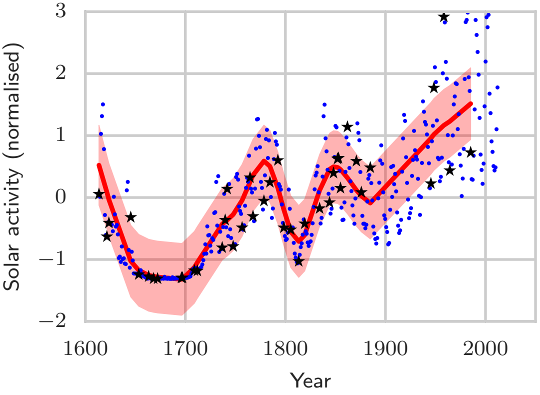
$\theta_1 \sim \mathcal{N}(0, 10); \theta_2 \sim \mathcal{N}(0, 1)$
$X_i \sim \frac{1}{2}\mathcal{N}(\theta_1, 4) + \frac{1}{2}\mathcal{N}(\theta_1 + \theta_2, 4)$.

Approximate MCMC: tradeoff between bias and computation (e.g. *Austerity in MCMC Land* [2])



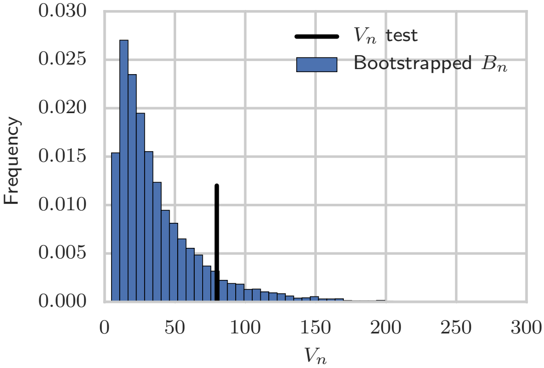
$\theta_1 \sim \mathcal{N}(0, 10); \theta_2 \sim \mathcal{N}(0, 1)$
$X_i \sim \frac{1}{2}\mathcal{N}(\theta_1, 4) + \frac{1}{2}\mathcal{N}(\theta_1 + \theta_2, 4)$.

Experiment 2: Statistical model criticism



We test the hypothesis that a Gaussian process model, learned from **training data** ⋆, is a good fit for the test data [3].

Experiment 2: Statistical model criticism



We test the hypothesis that a Gaussian process model, learned from **training data** $\star$, is a good fit for the test data [3].

Questions?

References

[1] J. Gorham and L. Mackey.
Measuring sample quality with stein's method.
In *NIPS*, pages 226–234, 2015.

[2] Anoop Korattikara, Yutian Chen, and Max Welling.
Austerity in mcmc land: Cutting the metropolis-hastings budget.
arXiv preprint arXiv:1304.5299, 2013.

[3] James R Lloyd and Zoubin Ghahramani.
Statistical model criticism using kernel two sample tests.
In *NIPS*, pages 829–837, 2015.

[4] C. Oates, M. Girolami, and N. Chopin.
Control functionals for monte carlo integration, 2015.

Stein's trick in the RKHS

Consider the class

$$G = \{f' + \log' p \cdot f | f \in \mathcal{F}\}$$

Stein's trick in the RKHS

Consider the class

$$G = \{f' + \log' p \cdot f | f \in \mathcal{F}\}$$

Given $g \in G$, then (integration by parts)

$$\begin{aligned}
\mathbb{E}_p g(X) &= \mathbb{E}_p \left[f'(X) + \log' p(X) f(X) \right] \\
&= \int f(x)' p(x) + f(x) p'(x) dx \\
&= \int_{-\infty}^{\infty} (f(x) p(x))' dx \\
&= f(x) p(x) \Big|_{x=-\infty}^{x=\infty} \\
&= 0
\end{aligned}$$

See [1, 4].