HARVARD UNIVERSITY

Microsoft

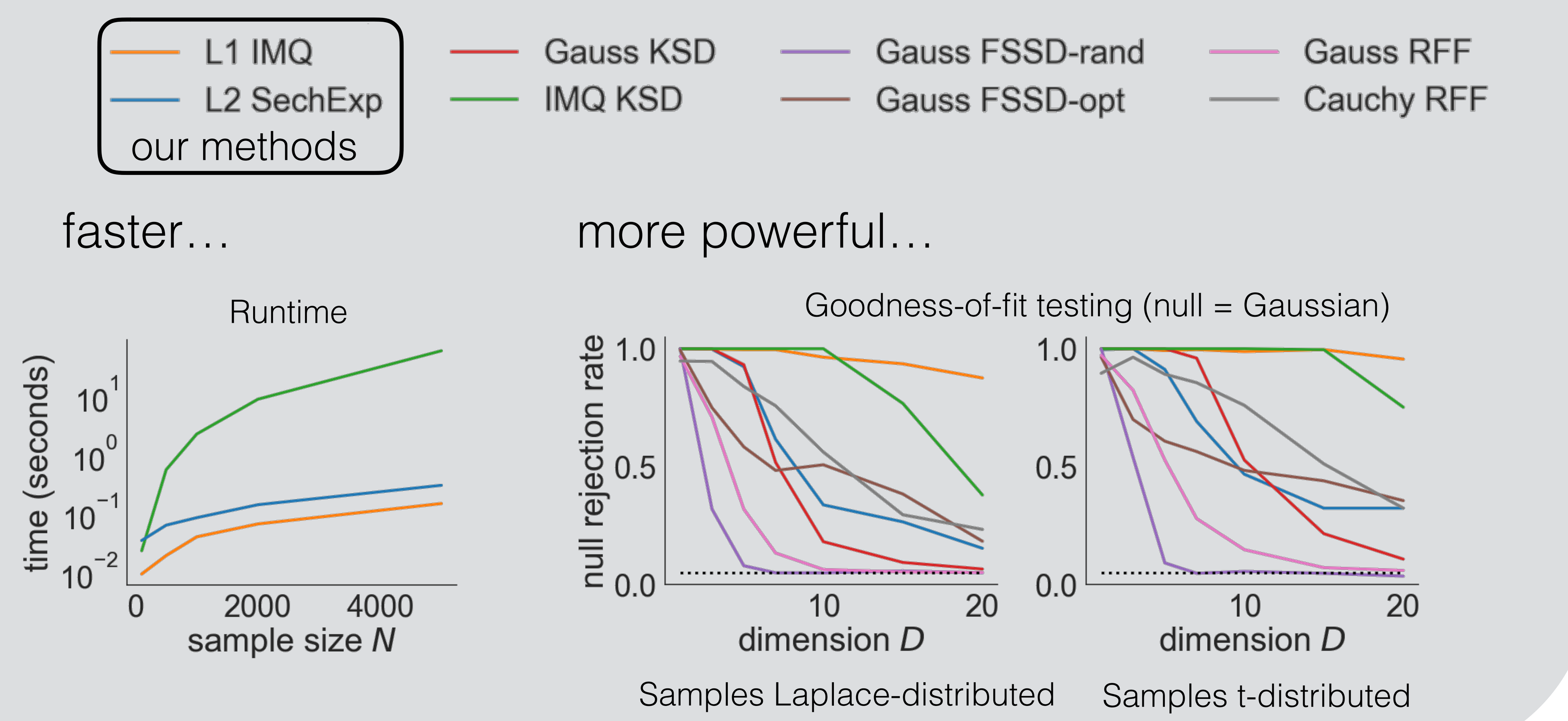
Random Feature Stein Discrepancies

Jonathan Huggins
Department of Biostatistics, Harvard University

Lester Mackey
Microsoft Research New England

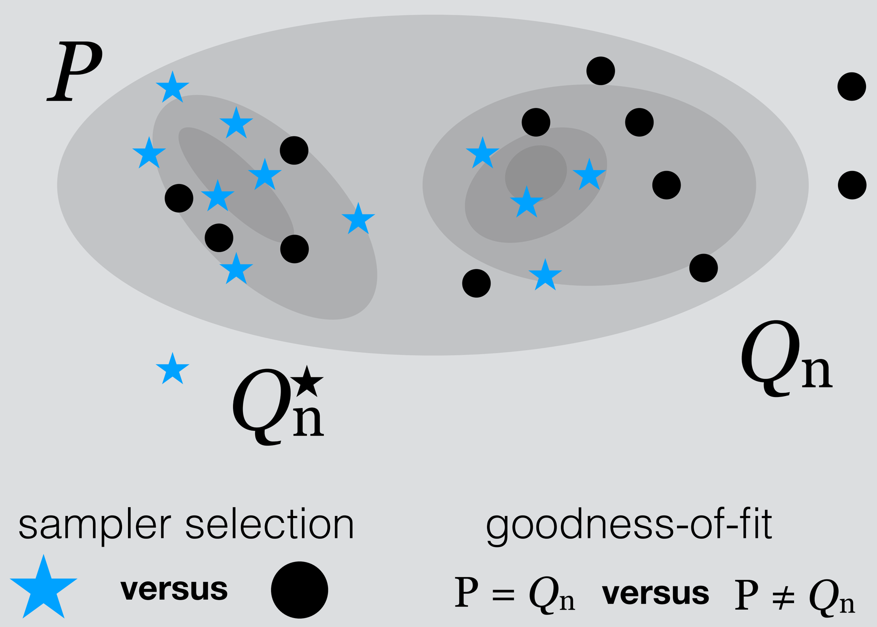
Summary

- **Computable Stein discrepancies** used for…
  - sampler selection
  - posterior inference
  - goodness-of-fit testing
- **But** computation scales **quadratically** with sample size
- We introduce **random feature Stein discrepancies**, which…
  - retain excellent theoretical properties of existing Stein discrepancies
  - are computable in **linear time**



The Big Picture

I. Motivation



II. Stein discrepancies

$$d_{\mathcal{T}\mathcal{G}}(Q_n, P) = \sup_{g \in \mathcal{G}} |Q_n(\mathcal{T}g) - \cancel{P(\mathcal{T}g)}|$$

Stein operator: $P(\mathcal{T}g) = 0$

III. Desiderata

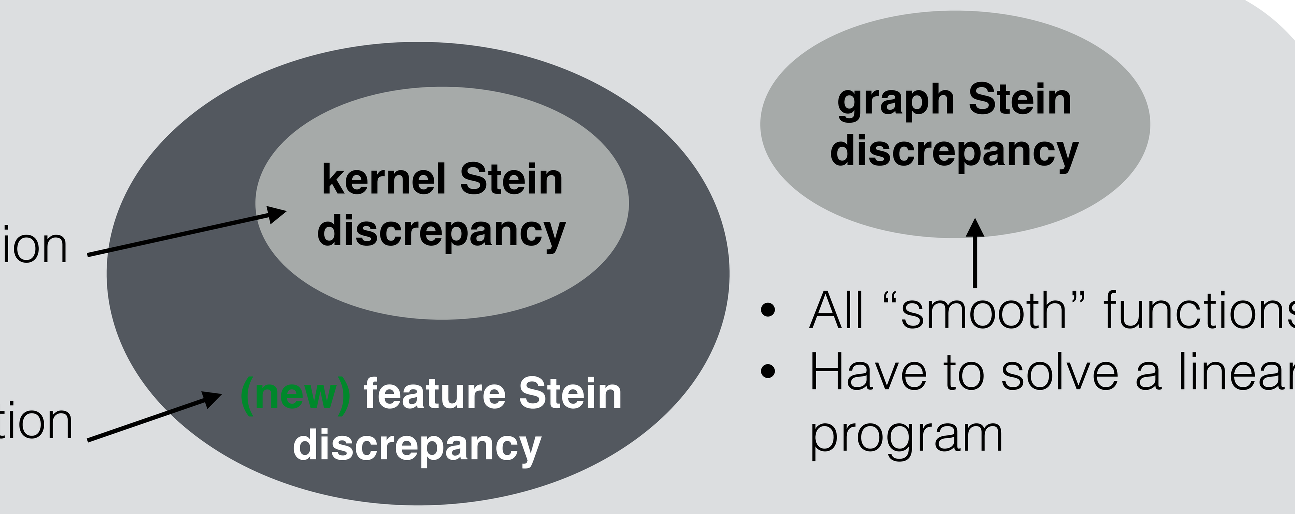
1. Detect convergence of Q_n to P
2. Detect non-convergence
3. Computationally efficient

IV. Choices for $\mathcal{G}$

- Uses a kernel function
- Closed-form

- Uses a feature function
- Approximate with importance sampling

- All "smooth" functions
- Have to solve a linear program

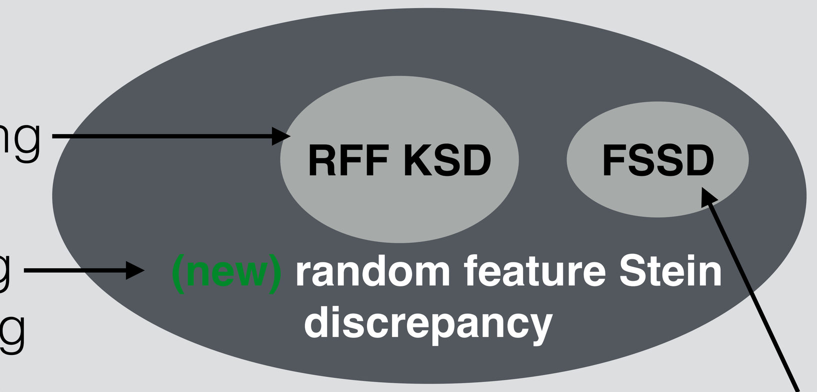


V. Linear-time approximations

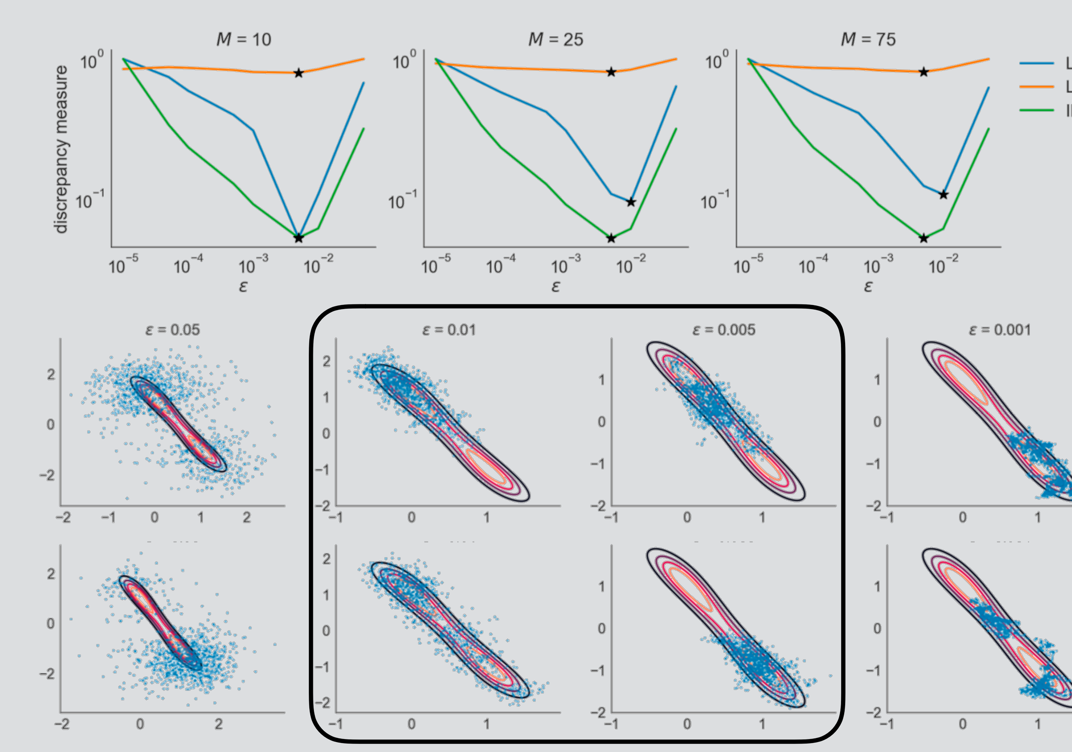
- Not convergence-determining

- Can be convergence-determining
- Provably close to ΦSD when using near-linear number of importance samples [$\Omega(n^\kappa)$ for any $\kappa > 0$]

- Underpowered



VI. Experiment: Selecting SGLD step size



10 importance samples suffices

ε = 0.01 or 0.005 gives highest-quality samples

All the Details

I. Defining the (random) feature Stein discrepancy

$$\mathcal{G}_{\Phi,r} \triangleq \left\{ g : \mathbb{R}^D \to \mathbb{R} \,|\, g_d(x) = \int \Phi(x,z)\overline{f_d(z)}\,\mathrm{d}z \text{ with } \textstyle\sum_{d=1}^D \|f_d\|^2_{L^s} \le 1 \text{ for } s = \frac{r}{r-1} \right\}.$$

Feature Stein discrepancy: $\Phi\mathrm{SD}^2_{\Phi,r}(\mu, P) \triangleq \sup_{g \in \mathcal{G}_{\Phi,r}} |\mu(\mathcal{T}g)|^2 = \sum_{d=1}^D \|\mu(\mathcal{T}_d\Phi)\|^2_{L^r}$

Random feature Stein discrepancy: for $Z_1, \dots, Z_M \overset{\text{i.i.d.}}{\sim} \nu$,

$$\mathrm{R\Phi SD}^2_{\Phi,r,\nu,M}(\mu, P) \triangleq \sum_{d=1}^D \left(M^{-1} \sum_{m=1}^M \nu(Z_m)^{-1} |\mu(\mathcal{T}_d\Phi)(Z_m)|^r \right)^{2/r}$$

II. Assumed form for feature function

Assumption: The base kernel has the form $k(x,y) = A_n(x)\Psi(x-y)A_n(y)$ for $A_n(x) \triangleq A(x - m_n)$ and $m_n \triangleq \mathbb{E}_{X \sim Q_n}[X]$

Assumption: $\Phi(x,z) = A_n(x)F(x-z)$

III. Detecting convergence and non-convergence

Proposition. *If the* tilted Wasserstein distance

$$\mathcal{W}_{A_n}(Q_n, P) \triangleq \sup_{h \in \mathcal{H}} |Q_n(A_n h) - P(A_n h)| \quad (\mathcal{H} \triangleq \{h : \|\nabla h(x)\|_2 \le 1, \forall x \in \mathbb{R}^D\})$$

converges to zero, then $\Phi\mathrm{SD}_{\Phi,r}(Q_n, P) \to 0$ *and* $\mathrm{R\Phi SD}_{\Phi,r,\nu_n,M_n}(Q_n, P) \xrightarrow{P} 0$ *for any choices of* $r \in [1,2]$, ν_n, *and* $M_n \ge 1$.

Proposition (KSD-ΦSD inequality)**.** *If* $k(x,y) = \int \mathcal{F}(\Phi(x,\cdot))(\omega)\overline{\mathcal{F}(\Phi(y,\cdot))(\omega)}\rho(\omega)\,\mathrm{d}\omega$, $r \in [1,2]$, *and* $\rho \in L^t$ *for* $t = r/(2-r)$, *then*

$$\mathrm{KSD}^2_k(Q_n, P) \le \|\rho\|_{L^t}\, \Phi\mathrm{SD}^2_{\Phi,r}(Q_n, P).$$

IV. Constructing convergence-determining RΦSDs

(C,γ) **second moments:** We say (Φ, r, ν) *yields* (C,γ) *second moments for* P *and* Q_n if $\mathbb{E}[Y^2_{n,d}] \le C\mathbb{E}[Y_{n,d}]^{2-\gamma}$ for $Y_{n,d} \triangleq |(Q_n\mathcal{T}_d\Phi)(Z)|^r/\nu(Z)$ and $Z \sim \nu$.

Proposition. *Suppose* (Φ, r, ν) *yields* (C,γ) *second moments for* P *and* Q_n. *If the reference* $\mathrm{KSD}_k(Q_n, P) = \Omega(n^{-1/2})$ *then a sample size* $M = \Omega(n^{\gamma r/2})$ *suffices to have, with high probability,*

$$2\|\rho\|^{1/2}_{L^t}\, \mathrm{R\Phi SD}_{\Phi,r,\nu,M}(Q_n, P) \ge \mathrm{KSD}_k(Q_n, P).$$

Theorem. *Under regularity conditions, there exists a* smoothness parameter $\overline{\lambda} \in (1/2, 1]$ *and a constant* $b \in [0,1)$ *such that the following holds. For any* $\xi \in (0, 1-b)$, $c > 0$, *and* $\alpha > 2(1 - \overline{\lambda})$, *if* $\nu(z) \ge c\,\Psi(z - m_N)^{\xi r}$, *then there exists a constant* $C_\alpha > 0$ *such that* (Φ, r, ν) *yields* $(C_\alpha, \gamma_\alpha)$ *second moments for* P *and* Q_n, *where* $\gamma_\alpha \triangleq \alpha + (2-\alpha)\xi/(2-b-\xi)$.

Tilted hyperbolic secant kernel:

$$\Psi(x) = \Psi^{\mathrm{sech}}_a(x) \triangleq \prod_{d=1}^D \mathrm{sech}\left(\sqrt{\tfrac{\pi}{2}}\,a x_d\right) \quad \text{and} \quad A(x) = \prod_{d=1}^D e^{c\sqrt{1+x_d^2}}$$

L^2 **tilted hyperbolic secant** RΦSD**:**

$$F = \Psi^{\mathrm{sech}}_{2a}, \quad r = 2, \quad \text{and} \quad \nu(z) \propto \Psi^{\mathrm{sech}}_{4a\xi}(z - m_n)$$

Inverse multiquadric kernel: $\Psi^{\mathrm{IMQ}}_{c,\beta}(x) \triangleq (c^2 + \|x\|_2^2)^\beta$ for some $\beta < 0$

L^r **IMQ** RΦSD**:**

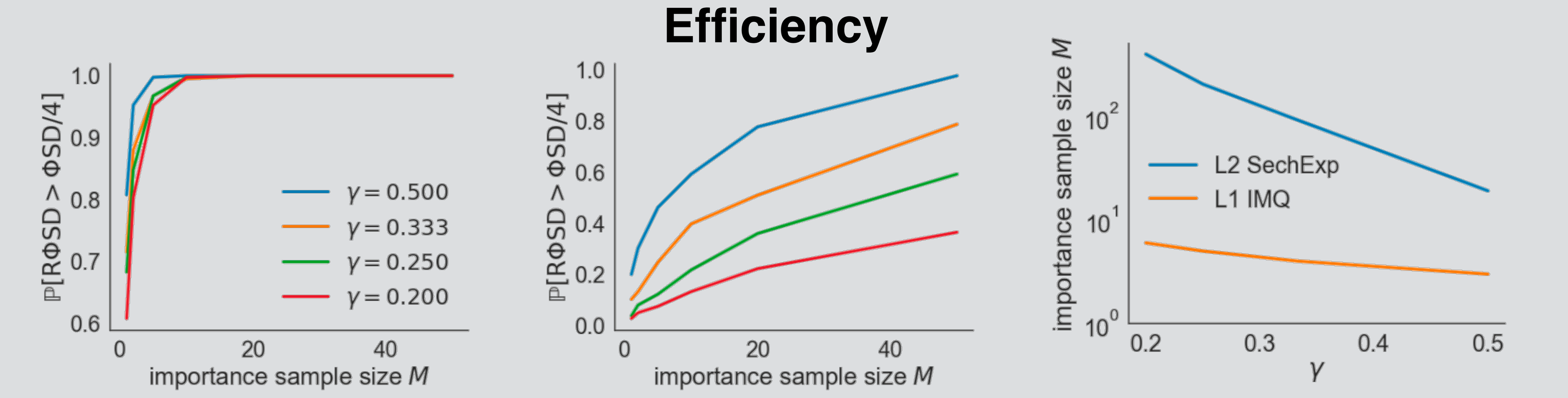
$$F = \Psi^{\mathrm{IMQ}}_{c',\beta'}, \quad r = -D/(2\beta'\underline{\xi}), \quad \text{and} \quad \nu(z) \propto \Psi^{\mathrm{IMQ}}_{c',\beta'}(z - m_N)^{\xi r},$$

where $c' = \overline{\lambda}c/2$, $\beta' \in [-D/(2\underline{\xi}), -\beta/(2\underline{\xi}) - D/(2\underline{\xi}))$, and $\xi \in (\underline{\xi}, 1)$

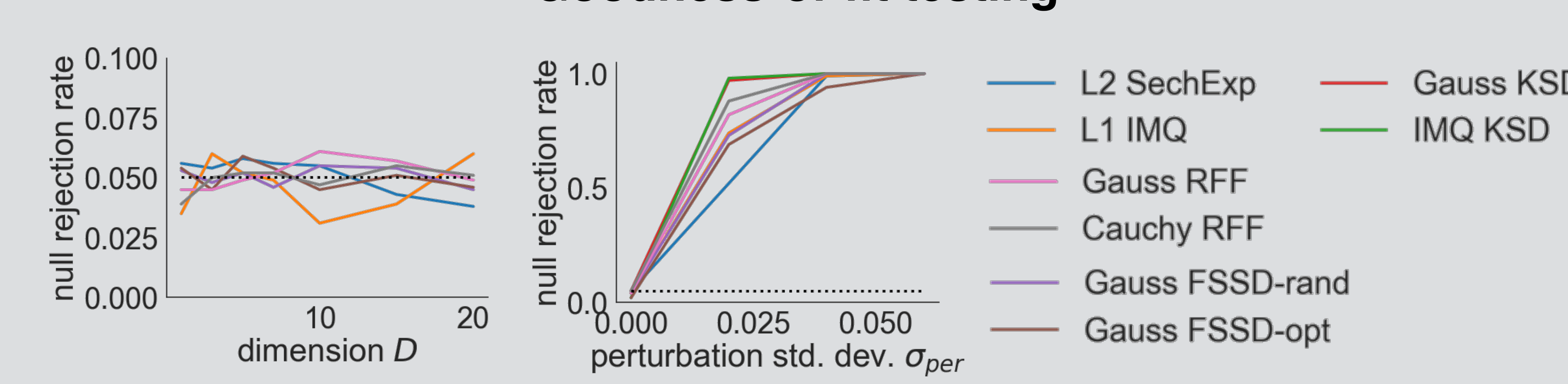
Simplest choice: $\beta' = -D/(2\underline{\xi})$ yields $r = 1$.

More Experiments

Efficiency



Goodness-of-fit testing



jhuggins@mit.edu • lmackey@microsoft.com

https://bitbucket.org/jhhuggins/random-feature-stein-discrepancies