We thank all reviewers for their encouraging and constructive feedback and respond to each in turn.

(R1) SSD advantages: The chief advantages of SSDs are that they require orders of magnitude less computation than SDs (while still determining convergence), can be deployed when exact SDs are simply infeasible (e.g., in the settings motivating many approximate MCMC methods), and, for a fixed computational budget, typically yield more accurate posterior approximations than exact SVGD.

(R1) Gaussian kernel: Thm. 6 of [22] showed that (non-stochastic) KSDs based on the Gaussian kernel fail to detect non-convergence (and thus often have terrible power in practice) even for simple target distributions like multivariate Gaussians; for the same reason SSDs based on the same kernel fail to detect non-convergence. We focused on the inverse multiquadric kernel (which has very different properties from a polynomial kernel), because its KSD detects non-convergence in great generality. Our results apply to other kernels, such as the inverse log kernel of [9, Thm. 3].

(R1) Thm. 4: We apologize for the confusion: the text preceding Thm. 4 is the contrapositive of (and hence equivalent to) the statement in Thm. 4. We will reword Thm. 4 to improve clarity and note here that Thm. 4 does not prove that $\mathcal{S}(\nu_n) \to 0$ but rather that $\nu_n \Rightarrow P$ if $\mathcal{S}(\nu_n) \to 0$ and ν_n is tight. We will also provide a roadmap at the start of Sec. 4.2 to clarify how the results fit together: we show that SSDs detect non-convergence (Thm. 6) in a series of steps: (a) by Thm. 4, if $Q_n \not\Rightarrow P$ then either a bounded SD $\not\to 0$ or Q_n is not tight; (b) by Thm. 3, if a bounded SD $\not\to 0$ then its SSD $\not\to 0$ w.p. 1; (c) by Prop. 5, if Q_n is not tight, then the SSD $\not\to 0$ surely. Here, the new result on bounded SD non-convergence (Thm. 4) is an important stepping stone to establishing SSD non-convergence (Thm. 6).

(R2) Minibatches: In the revision, we will clarify that a separate minibatch per sample point is standard in the SD context: it is used in the original SVGD paper [32] and each of the cited uses of SSDs [2, 40]. A separate minibatch per sample point is also standard in each of the approximate MCMC algorithms discussed [8, 14], including stochastic gradient Langevin dynamics [48] and SGFS [1]. We will also highlight the substantial advantage of using separate minibatches over a single minibatch. If P is the target and $\tilde{P}$ is the posterior induced by a single minibatch of data, then the separate minibatch SSD is guaranteed to detect convergence and non-convergence to P for any minibatch size (by our Thms. 2 & 6), but a single minibatch SSD cannot correctly discriminate between P and $\tilde{P}$ (it will incorrectly declare that samples from $\tilde{P}$ are converging to P and incorrectly declare that samples from P are not converging to P).

(R2) App. E: Thank you for pointing out this inadvertent omission. The revision will reflect that, exactly as in the proof of Thm. 4, the other two cases follow as their Stein sets contain a scaled copy of the kernel Stein set.

(R2) App. F: Thank you for flagging this error. We have corrected the statement using $\mathcal{H} = \{h : \|h\|_\infty + \mathrm{Lip}(h) < 1\}$:

Lemma 1. *If two sequences of random measures $(\nu_n)_{n=1}^\infty$ and $(\tilde{\nu}_n)_{n=1}^\infty$ on $\mathbb{R}^d$ satisfy $\nu_n(hI_{B_R}) - \tilde{\nu}_n(hI_{B_R}) \overset{a.s.}{\to} 0$ for each $h \in C_b$ and some $B_R \triangleq \{\|x\|_2 \le R\}$ with $R \ge S$ for all $S > 0$, then $\sup_{h\in\mathcal{H}} |\nu_n(hI_{B_R}) - \tilde{\nu}_n(hI_{B_R})| \overset{a.s.}{\to} 0$ for each $R > 0$. If, in addition, f_0 is almost surely uniformly ν_n-integrable and uniformly $\tilde{\nu}_n$-integrable, and f_0, f_1 are bounded on compact sets, then $\sup_{h\in\mathcal{H}_f} |\nu_n(h) - \tilde{\nu}_n(h)| \overset{a.s.}{\to} 0$ for $\mathcal{H}_f = \{h : |h| \le f_0, \frac{|h(x)-h(y)|}{\|x-y\|_2} \le f_1(x), \forall x, y\}$.*

Proof Fix $R, \epsilon > 0$, and let $K = B_R$. By the Arzelà–Ascoli theorem, there exists a finite $\epsilon/2$ subcover of the set of K-restrictions $\{h|_K : h \in \mathcal{H}\}$, extendable to C_b functions $(h_k)_{k=1}^m$ on $\mathbb{R}^d$. The union bound and our assumption now give $\mathbb{P}(\sup_{h\in\mathcal{H}} |\nu_n(hI_K) - \tilde{\nu}_n(hI_K)| > \epsilon \text{ i.o.}) \le \mathbb{P}(\max_{1\le k\le m} |\nu_n(h_kI_K) - \tilde{\nu}_n(h_kI_K)| > \epsilon/2 \text{ i.o.})$ $\le \sum_{k=1}^m \mathbb{P}(|\nu_n(h_kI_K) - \tilde{\nu}_n(h_kI_K)| > \epsilon/2 \text{ i.o.}) = 0$. As $\mathcal{H}_f$ is uniformly bounded-Lipschitz on B_R, the second claim follows as in the submission with $\mathcal{H}_f$ and $\mathcal{H}$ replacing $C(\mathbb{R}^d) : |h| \le |f|$ and $C(\mathbb{R}^d)$ and $K_\epsilon = B_R$ for suitable R. □

For any target P, there exists a sequence of radii $(R_j)_{j=1}^\infty$ with $R_j \to \infty$ such that $\mathbb{E}_P(\|X\|_2 = R_j) = 0$ so that B_{R_j} is a continuity set under P. Since $Q_n \Rightarrow P$, we have $Q_n(hI_{B_{R_j}}) \to P(hI_{B_{R_j}})$ for each j and $h \in C_b$ by the Portmanteau theorem. Thm. 2 now follows assuming $\sup_{g\in\mathcal{G}_n,y} \|\mathcal{T}_l g(x) - \mathcal{T}_l g(y)\|_2/\|x-y\|_2$ bounded on compact sets (which holds for the pairings in [21-23]), and Thm. 7 and Lem. 11 follow assuming $\sup_{l\in[L],z\in\mathbb{R}^d} \|\nabla_x(\nabla \log p_l(x)k(x,z))\|_2$ bounded on compact sets (which holds for $\nabla \log p_l$ in C^1 and bounded-Lipschitz k).

(R3) Subsampling: Prior work, like the finite set SD of [28] and the random feature SDs of [27] address the n^2 complexity of SDs by introducing alternative SDs with O(n) complexity. We will clarify that our work addresses the complementary problem of an expensive Stein operator and should not be viewed as an alternative to O(n)-time SDs. Rather, datapoint subsampling can be directly applied to O(n) SDs to obtain O(n) SSDs with additional speed-ups.

(R3) Discrete: In the revision, we will clarify that while we develop the most extensive theory for the popular Langevin Stein operator, our results on detecting convergence (Thm. 2), enforcing tightness (Prop. 5), and detecting bounded non-convergence (Thm. 3) apply to any Stein operator and to both discrete and continuous targets.

(R3) Coordinate kernels: Thank you for these references. In the revision, we will highlight that both works can be viewed as deploying exact SDs with special Stein sets featuring coordinate-dependent kernels. Since every coordinate is still updated on each SVGD step, this is somewhat different from, for example, subsampling coordinate operators for computational benefit (in which case certain coordinates would not be updated at all on each SVGD step). However, these SDs can be combined with datapoint subsampling to obtain substantial speed-ups.