SUPPLEMENTARY MATERIALS FOR "A VARIANCE REDUCTION METHOD FOR NEURAL-BASED DIVERGENCE ESTIMATION"

Anonymous authors
Paper under double-blind review

A NOTATION

We denote the extended reals by $\overline{\mathbb{R}} \equiv \mathbb{R} \cup \{-\infty, \infty\}$. We let $(\Omega, \mathcal{M})$ be a measurable space and $\mathcal{P}(\Omega)$ be the set of probability measures on $(\Omega, \mathcal{M})$. For $P \in \mathcal{P}(\Omega)$ we let E_P denote the expectation with respect to P and Var_P denote the variance with respect to P. We let Q_n, P_n be n-sample empirical measures, constructed using i.i.d. samples from Q and P respectively. A set $\Psi \subset \mathcal{P}(\Omega)$ will be called $\mathcal{P}(\Omega)$**-determining** if whenever $Q, P \in \mathcal{P}(\Omega)$ satisfy $E_Q[g] = E_P[g]$ for all $g \in \Psi$ we have $Q = P$. A map $D : \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \to [0, \infty]$ will be said to have the **divergence property** if $D(Q\|P) = 0$ if and only if $Q = P$.

B DIVERGENCE PROPERTY

In this appendix we prove the divergence property for the variance-penalized divergences introduced in Section 3.1 (the definition of the divergence property is found in Appendix A). The techniques we use are based on the proof of Theorem C.3 in Birrell et al. (2020). Though we focus on the Rényi, KL, and f-divergences the same techniques can be adapted to other divergences with a variational formulation, such as integral probability metrics.

Theorem 1 (Divergence Property for $R_\alpha^{\Gamma,\lambda}$)**.** *Let $\alpha > 0$, $\alpha \neq 1$, $\Gamma \subset \mathcal{M}_b(\Omega)$, $\lambda \geq 0$, and suppose there exists a nonempty set $\Psi \subset \mathcal{M}_b(\Omega)$ such that*

1. *Ψ is $\mathcal{P}(\Omega)$-determining,*
2. *for all $\psi \in \Psi$ there exists $\epsilon_0 > 0$ with $\epsilon\psi \in \Gamma$ for all $|\epsilon| < \epsilon_0$.*

Then $R_\alpha^{\Gamma,\lambda}$ has the divergence property.

Remark 2. *The definition of $\mathcal{P}(\Omega)$-determining can be found in Appendix A.*

Proof. Let $Q, P \in \mathcal{P}(\Omega)$. The second assumption on Ψ implies $0 \in \Gamma$ so we can bound equation (15) below by the value at $g = 0$, which implies $R_\alpha^{\Gamma,\lambda}(Q\|P) \geq 0$. The variance penalty is non-negative and so we have

$$0 \leq R_\alpha^{\Gamma,\lambda}(Q\|P) \leq R_\alpha^{\Gamma}(Q\|P) \leq R_\alpha(Q\|P)\,, \tag{1}$$

where we used equation (1). R_α is known to satisfy the divergence property and so if $Q = P$ then $R_\alpha(Q\|P) = 0$, hence $R_\alpha^{\Gamma,\lambda}(Q\|P) = 0$ as well.

Finally, suppose $R_\alpha^{\Gamma,\lambda}(Q\|P) = 0$. Given $\psi \in \Psi$ let $g_\epsilon = \epsilon\psi \in \Gamma$ for $|\epsilon| < \epsilon_0$ (here we are using the second assumption on Ψ). Then we can bound equation (15) below as follows

$$\begin{aligned} 0 =& R_\alpha^{\Gamma,\lambda}(Q\|P) \\ \geq& \frac{1}{\alpha - 1} \log E_Q[e^{(\alpha-1)\epsilon\psi}] - \frac{1}{\alpha} \log E_P\left[e^{\alpha\epsilon\psi}\right] \\ & - \lambda \left(\frac{1}{(\alpha-1)^2} \frac{\mathrm{Var}_Q[e^{(\alpha-1)\epsilon\psi}]}{(E_Q[e^{(\alpha-1)\epsilon\psi}])^2} + \frac{1}{\alpha^2} \frac{\mathrm{Var}_P[e^{\alpha\epsilon\psi}]}{(E_P[e^{\alpha\epsilon\psi}])^2} \right) \equiv h(\epsilon)\,. \end{aligned} \tag{2}$$

It is straightforward to calculate $h(0) = 0$ and $h'(0) = E_Q[\psi] - E_P[\psi]$ (in particular, the first derivative of the variance penalty vanishes at $\epsilon = 0$). These, together with the fact that $h(\epsilon)$ is C^1 and $h(\epsilon) \leq 0$ for all $|\epsilon| < \epsilon_0$ imply

$$0 = h'(0) = E_Q[\psi] - E_P[\psi]\,. \tag{3}$$

This holds for all $\psi \in \Psi$ and Ψ was assumed to be $\mathcal{P}(\Omega)$-determining, hence we conclude that $Q = P$. This completes the proof of the divergence property for $R_\alpha^{\Gamma,\lambda}$. □

Theorem 3 (Divergence Property for $D_{\mathrm{KL}}^{\Gamma,\lambda}$)**.** *Let $\Gamma \subset \mathcal{M}_b(\Omega)$, $\lambda \geq 0$, and suppose there exists a nonempty set $\Psi \subset \mathcal{M}_b(\Omega)$ such that*

1. *Ψ is $\mathcal{P}(\Omega)$-determining,*
2. *for all $\psi \in \Psi$ there exists $\epsilon_0 > 0$ with $\epsilon\psi \in \Gamma$ for all $|\epsilon| < \epsilon_0$.*

Then $D_{KL}^{\Gamma,\lambda}$ has the divergence property.

The proof of Theorem 3 follows the same template as the proof of Theorem 1; we omit the details. The proof for $D_f^{\Gamma,\lambda}$ divergences is slightly more nuanced than the Rényi or KL cases, and so we provide a proof below.

Theorem 4 (Divergence property for $D_f^{\Gamma,\lambda}$)**.** *Let $\lambda > 0$, $\Gamma \subset \mathcal{M}_b(\Omega)$, and $f : \mathbb{R} \to (-\infty, \infty]$ be lower semicontinuous and convex, with $f(1) = 0$. Suppose f and Γ also satisfy the following:*

1. *There exist a nonempty set $\Psi \subset \Gamma$ with the following properties:*
   (a) *Ψ is $\mathcal{P}(\Omega)$-determining.*
   (b) *For all $\psi \in \Psi$ there exists $c_0 \in \mathbb{R}$, $\epsilon_0 > 0$ such that $c_0 + \epsilon\psi \in \Gamma$ for all $|\epsilon| < \epsilon_0$.*
2. *f is finite and strictly convex on a neighborhood of 1.*
3. *f^* is finite and C^1 on a neighborhood of $\nu_0 \equiv f'_+(1)$.*

Then $D_f^{\Gamma,\lambda}$ has the divergence property.

Remark 5. *f'_+ denotes the right derivative.*

Proof. The assumptions on f imply that $f^*(\nu_0) = \nu_0$ and $(f^*)'(\nu_0) = 1$ (see Lemma A.9 in Birrell et al. (2020)). Assumption 1 implies there exists $c_0 \in \Gamma \cap \mathbb{R}$, hence we can bound equation (12) below by its value at $g = c_0$, $\nu = c_0 - \nu_0$ to find

$$D_f^{\Gamma,\lambda}(Q\|P) \geq E_Q[\nu_0] - E_P[f^*(\nu_0)] = 0 \tag{4}$$

(note that the variance penalty vanishes here). We clearly have $D_f^{\Gamma,\lambda} \leq D_f^\Gamma \leq D_f$. $D_f(Q\|P) = 0$ when $Q = P$, hence $D_f^{\Gamma,\lambda}(Q\|P) = 0$ as well.

Finally, suppose $D_f^{\Gamma,\lambda}(Q\|P) = 0$. From assumption 1.b, given $\psi \in \Psi$ there exists $c_0 \in \mathbb{R}$, $\epsilon_0 > 0$ such that $g_\epsilon \equiv c_0 + \epsilon\psi \in \Gamma$ for all $|\epsilon| < \epsilon_0$. Therefore

$$\begin{aligned} 0 = D_f^{\Gamma,\lambda}(Q\|P) \geq & E_Q[g_\epsilon - (c_0 - \nu_0)] - E_P[f^*(g_\epsilon - (c_0 - \nu_0))] \\ & - \lambda(\mathrm{Var}_Q[g_\epsilon - (c_0 - \nu_0)] + \mathrm{Var}_P[f^*(g_\epsilon - (c_0 - \nu_0))]) \\ = & \nu_0 + \epsilon E_Q[\psi] - E_P[f^*(\nu_0 + \epsilon\psi)] - \lambda(\mathrm{Var}_Q[\nu_0 + \epsilon\psi] + \mathrm{Var}_P[f^*(\nu_0 + \epsilon\psi)]) \\ \equiv & h(\epsilon)\,. \end{aligned} \tag{5}$$

As computed in equation (4), we have $h(0) = 0$. Together with the fact that $h(\epsilon)$ is C^1 and $h(\epsilon) \leq 0$ for all $|\epsilon| < \epsilon_0$ we can conclude that

$$0 = h'(0) = E_Q[\psi] - E_P[(f^*)'(\nu_0)\psi] = E_Q[\psi] - E_P[\psi] \tag{6}$$

(again, the first derivative of the variance penalty vanishes at $\epsilon = 0$). Equation (6) holds for all $\psi \in \Psi$, a $\mathcal{P}(\Omega)$-determining set. Hence $Q = P$. This completes the proof of the divergence property for $D_f^{\Gamma,\lambda}$. □

We end this appendix by presenting several examples of $\mathcal{P}(\Omega)$-determining sets that can be used to construct Γ's satisfying the assumptions of the above theorems.

1. Exponentials, $e^{c\cdot x}$, $c \in \mathbb{R}^n$, i.e., the moment generating function; see Section 30 in Billingsley (2012).
2. The set of bounded continuous functions on a metric space.
3. The set of continuous functions on a metric space with $\|g\|_\infty \leq 1$ ($\|\cdot\|_\infty$ denotes the supremum norm).
4. The set of bounded 1-Lipschitz functions on a metric space.
5. The set of bounded 1-Lipschitz functions, g, on a metric space with $\|g\|_\infty \leq 1$ (items 2-5 follow from Theorem 2.1 in Billingsley (2013)).
6. The unit ball in a reproducing kernel Hilbert space, under appropriate assumptions (see Sriperumbudur et al. (2011)).
7. The set of ReLU neural networks. This follows from the universal approximation theorem (Cybenko, 1989) and also applies to other activation functions, e.g., sigmoid.
8. The set of ReLU neural networks with spectral normalization (Miyato et al., 2018).

C Convergence Proofs

In this appendix we consider the limit of the variance-penalized divergences as the penalty strength approaches 0 or ∞. First we consider the easier case where the penalty strength approaches zero. For this we will use the following general convergence result.

Lemma 6. *Let $H[g] \in \mathbb{R}$, $V[g] \in [0,\infty)$ where $g \in \Gamma$ is an arbitrary index set. Define $D = \sup_{g\in\Gamma} H[g]$ and for $\lambda > 0$ define $D^\lambda = \sup_{g\in\Gamma}\{H[g] - \lambda V[g]\}$. Then*

$$\lim_{\lambda\to 0^+} D^\lambda = D. \tag{7}$$

Proof. The $V[g]$ are non-negative, hence D^λ is non-increasing in λ. Therefore

$$\lim_{\lambda\to 0^+} D^\lambda = \sup_{\lambda>0} D^\lambda = \sup_{\lambda>0}\sup_{g\in\Gamma}\{H[g] - \lambda V[g]\} = \sup_{g\in\Gamma}\sup_{\lambda>0}\{H[g] - \lambda V[g]\} = D\,. \tag{8}$$

□

The objective functionals of $D_{\mathrm{KL}}^{\Gamma,\lambda}$ and $R_\alpha^{\Gamma,\lambda}$ satisfy the assumptions of Lemma 6, as does $D_f^{\Gamma,\lambda}$'s if $f^*(y) < \infty$ for all $y \in \mathbb{R}$. Therefore Theorem 2 follows as a corollary to Lemma 6.

Next we investigate the limit as $\lambda \to \infty$. The proofs in this case are more involved than the $\lambda \to 0^+$ limit. We start with the following lemma.

Lemma 7. *Let $H[g] \in \mathbb{R}$, $V[g] \in [0,\infty)$ where $g \in \Gamma$ is an arbitrary index set. Define $D = \sup_{g\in\Gamma} H[g]$ and for $\lambda > 0$ define $D^\lambda = \sup_{g\in\Gamma}\{H[g] - \lambda V[g]\}$. Suppose $D^\lambda \geq 0$ and $D < \infty$. Then for $n \in \mathbb{Z}^+$ there exists $g_n \in \Gamma$ with*

$$0 \leq D^n \leq H[g_n] + 1/n \;\; \textit{and} \;\; \lim_{n\to\infty} V[g_n] = 0\,. \tag{9}$$

Proof. We have $0 \leq D^\lambda \leq D < \infty$ and so the D^λ are finite. The definition of D^λ then implies that there exists g_n with

$$D^n - 1/n \leq H[g_n] - nV[g_n] \tag{10}$$

for all n. The assumptions $V[g_n] \geq 0$ and $D^n \geq 0$ then imply $0 \leq D^n \leq H[g_n] + 1/n$. The inequality (10) also implies

$$0 \leq V[g_n] \leq H[g_n]/n - D^n/n + 1/n^2 \leq H[g_n]/n + 1/n^2 \leq D/n + 1/n^2 \to 0 \tag{11}$$

as $n \to \infty$. Therefore $\lim_{n\to\infty} V[g_n] = 0$ as claimed. □

Next we consider the implications of $V[g_n] \to 0$ when V takes the form of one of the variance penalties from Section 3.1.

Lemma 8. *Let $P \in \mathcal{P}(\Omega)$ and $h_n \in \mathcal{M}_b(\Omega)$.*

1. *If*

$$\lim_{n\to\infty} \mathrm{Var}_P[h_n] = 0 \tag{12}$$

then there exists a subsequence h_{n_j} such that

$$h_{n_j} - E_P[h_{n_j}] \to 0 \;\; P\text{-a.s.} \tag{13}$$

2. *If $c \in \mathbb{R} \setminus \{0\}$ and*

$$\lim_{n\to\infty} \mathrm{Var}_P[e^{ch_n}]/(E_P[e^{ch_n}])^2 = 0 \tag{14}$$

then there exists a subsequence h_{n_j} such that

$$h_{n_j} - \frac{1}{c}\log E_P[e^{ch_{n_j}}] \to 0 \;\; P\text{-a.s.} \tag{15}$$

Proof. Equation (13) is a direct consequence of Corollary 2.32 in Folland (2013). As for the second claim, we can write

$$\mathrm{Var}_P[e^{ch_n}]/(E_P[e^{ch_n}])^2 = \mathrm{Var}_P[e^{ch_n}/E_P[e^{ch_n}]] \tag{16}$$

and so the first item implies there exists a subsequence with

$$e^{ch_{n_j}}/E_P[e^{ch_{n_j}}] - E_P[e^{ch_{n_j}}/E_P[e^{ch_{n_j}}]] \to 0 \;\; P\text{-a.s.}, \tag{17}$$

i.e.,

$$e^{ch_{n_j}}/E_P[e^{ch_{n_j}}] - 1 \to 0 \;\; P\text{-a.s.} \tag{18}$$

Adding 1 to both sides, taking the logarithm, and then dividing by c gives the claimed result.

□

We are now ready to prove the following limit results as $\lambda \to \infty$.

Theorem 9. *Let $Q, P \in \mathcal{P}(\Omega)$ with $Q \ll P$ and $\Gamma \subset \mathcal{M}_b(\Omega)$.*

1. *If $0 \in \Gamma$ and $D_{KL}^{\Gamma}(Q\|P) < \infty$ then $\lim_{\lambda\to\infty} D_{KL}^{\Gamma,\lambda}(Q\|P) = 0$.*
2. *Let $\alpha > 0$, $\alpha \neq 1$. If $0 \in \Gamma$ and $R_\alpha^{\Gamma}(Q\|P) < \infty$ then $\lim_{\lambda\to\infty} R_\alpha^{\Gamma,\lambda}(Q\|P) = 0$.*
3. *Let $f : \mathbb{R} \to (-\infty, \infty]$ be lower semicontinuous, convex, finite on a neighborhood of 1 with $f(1) = 0$, and satisfy $f^*(y) < \infty$ for all $y \in \mathbb{R}$. If there exists $c_0 \in \Gamma \cap \mathbb{R}$ and $D_f^{\Gamma}(Q\|P) < \infty$ then $\lim_{\lambda\to\infty} D_f^{\Gamma,\lambda}(Q\|P) = 0$.*

Remark 10. *Note that if $D_{KL}(Q\|P) < \infty$ then $D_{KL}^{\Gamma}(Q\|P) < \infty$ and similarly for the other divergences.*

Proof.

1. Let $D = D_{\mathrm{KL}}^{\Gamma}(Q\|P)$ and $D^\lambda = D_{\mathrm{KL}}^{\Gamma,\lambda}(Q\|P)$. The assumption $0 \in \Gamma$ allows us to bound equation (14) below by the value at $g = 0$, resulting in the bound $D^\lambda \geq 0$. This, together with equation (14) and the assumption that $D < \infty$, implies that D and D^λ are of the form required by Lemma 7, hence we conclude that there exists $g_n \in \Gamma$ with

$$0 \leq D^n \leq E_Q[g_n] - \log E_P[e^{g_n}] + 1/n \tag{19}$$

and

$$\lim_{n\to\infty} \left(\mathrm{Var}_Q[g_n] + \mathrm{Var}_P[e^{g_n}]/(E_P[e^{g_n}])^2\right) = 0\,. \tag{20}$$

Note that both terms in (20) are non-negative, hence they each converge to zero individually. Combining parts 1 and 2 of Lemma 8 we therefore obtain a subsequence g_{n_j} with

$$g_{n_j} - E_Q[g_{n_j}] \to 0 \;\; Q\text{-a.s.} \tag{21}$$

and

$$g_{n_j} - \log E_P[e^{g_{n_j}}] \to 0 \;\; P\text{-a.s.} \tag{22}$$

The assumption $Q \ll P$ implies that $g_{n_j} - \log E_P[e^{g_{n_j}}] \to 0$ Q-a.s. as well. Therefore there exists $\omega \in \Omega$ with $g_{n_j}(\omega) - E_Q[g_{n_j}] \to 0$ and $g_{n_j}(\omega) - \log E_P[e^{g_{n_j}}] \to 0$. Combining these facts with equation (19) we therefore obtain

$$0 \le D^{n_j} \le E_Q[g_{n_j}] - g_{n_j}(\omega) + g_{n_j}(\omega) - \log E_P[e^{g_{n_j}}] + 1/n_j \to 0 \tag{23}$$

as $j \to \infty$. Therefore $\lim_{j\to\infty} D^{n_j} = 0$. D^λ is non-increasing in λ, hence we can conclude that $\lim_{\lambda\to\infty} D^\lambda = 0$ as claimed.

2. The proof in this case is very similar to the proof of claim 1; we omit the details.

3. If we let $\nu_0 = f'_+(1)$ (right derivative) then $f^*(\nu_0) = \nu_0$ (see Lemma A.9 in Birrell et al. (2020)). Bounding equation (12) below by the value at $g = c_0$, $\nu = c_0 - \nu_0$ we see that $D_f^{\Gamma,\lambda}(Q\|P) \ge 0$. Similarly to the proof of part 1, using Lemmas 7 and 8 one can show that there exists $\omega \in \Omega$ and subsequences $g_{n_j} \in \Gamma$, $\nu_{n_j} \in \mathbb{R}$ such that

$$\begin{aligned} &g_{n_j}(\omega) - E_Q[g_{n_j}] \to 0\,, \\ &f^*(g_{n_j}(\omega) - \nu_{n_j}) - E_P[f^*(g_{n_j} - \nu_{n_j})] \to 0\,, \\ &0 \le D_f^{\Gamma,n_j}(Q\|P) \le \frac{1}{n_j} + E_Q[g_{n_j} - \nu_{n_j}] - E_P[f^*(g_{n_j} - \nu_{n_j})]\,. \end{aligned} \tag{24}$$

Combining these with the inequality

$$f^*(y) = \sup_{x\in\mathbb{R}}\{yx - f(x)\} \ge y - f(1) = y \tag{25}$$

we can conclude that $\lim_{j\to\infty} D_f^{\Gamma,n_j}(Q\|P) = 0$. $D_f^{\Gamma,\lambda}(Q\|P)$ is non-increasing in λ, hence this implies $\lim_{\lambda\to\infty} D_f^{\Gamma,\lambda}(Q\|P) = 0$ as claimed.

□

D BIAS BOUNDS

In this appendix we derive bounds on the bias of Rényi and f-divergence variational formula estimators. The key lemma is the following simple results regarding the expectation of a supremum or infimum.

Lemma 11. *Given an objective functional, $H : \mathcal{M}_b(\Omega) \times \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \to \overline{\mathbb{R}}$, and a test function space $\Gamma \subset \mathcal{M}_b(\Omega)$ we have*

$$\begin{aligned} \mathbb{E}[\sup_{g\in\Gamma} H[g; Q_n, P_n]] &\ge \sup_{g\in\Gamma} \mathbb{E}[H[g; Q_n, P_n]]\,, \\ \mathbb{E}[\inf_{g\in\Gamma} H[g; Q_n, P_n]] &\le \inf_{g\in\Gamma} \mathbb{E}[H[g; Q_n, P_n]]\,. \end{aligned} \tag{26}$$

The next lemma provides a bound on the bias of statistical estimators of Λ_f^P from equation (4).

Lemma 12. *Let f be convex with $f(1) = 0$, $P \in \mathcal{P}(\Omega)$, and P_n be n-sample empirical measures from Q and P respectively. Then for all $g \in \mathcal{M}_b(\Omega)$ the generalized cumulant generating function satisfies*

$$\mathbb{E}[\Lambda_f^{P_n}[g]] \le \Lambda_f^P[g]\,. \tag{27}$$

Proof. Using equation (4) we can compute

$$\begin{aligned}\mathbb{E}\left[\Lambda_f^{P_n}[g]\right] =& \mathbb{E}\left[\inf_{\nu\in\mathbb{R}}\{\nu + E_{P_n}[f^*(g-\nu)]\}\right] \\ \leq& \inf_{\nu\in\mathbb{R}}\mathbb{E}\left[\nu + E_{P_n}[f^*(g-\nu)]\right] \\ =& \inf_{\nu\in\mathbb{R}}\{\nu + E_P[f^*(g-\nu)]\} = \Lambda_f^P[g]\,.\end{aligned} \tag{28}$$

□

Lemmas 12 and 11 allow us to bound the bias of both f-divergences and Rényi divergences.

Corollary 13 (Rényi Divergence Bias Bound)**.** *For $\alpha \in (0,1)$ and $g \in \mathcal{M}_b(\Omega)$ we have*

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{\alpha-1}\log E_{Q_n}[e^{(\alpha-1)g}] - \frac{1}{\alpha}\log E_{P_n}[e^{\alpha g}]\right] \\ \geq& \frac{1}{\alpha-1}\log E_Q[e^{(\alpha-1)g}] - \frac{1}{\alpha}\log E_P[e^{\alpha g}]\end{aligned} \tag{29}$$

and

$$\mathbb{E}[R_\alpha^\Gamma(Q_n\|P_n)] \geq R_\alpha^\Gamma(Q\|P)\,. \tag{30}$$

Proof. To prove equation (29) we compute

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{\alpha-1}\log E_{Q_n}[e^{(\alpha-1)g}] - \frac{1}{\alpha}\log E_{P_n}[e^{\alpha g}]\right] \\ =& -\frac{1}{1-\alpha}\mathbb{E}\left[\Lambda^{Q_n}[(\alpha-1)g]\right] - \frac{1}{\alpha}\mathbb{E}\left[\Lambda^{P_n}[\alpha g]\right] \\ \geq& -\frac{1}{1-\alpha}\Lambda^Q[(\alpha-1)g] - \frac{1}{\alpha}\Lambda^P[\alpha g] \\ =& \frac{1}{\alpha-1}\log E_Q[e^{(\alpha-1)g}] - \frac{1}{\alpha}\log E_P[e^{\alpha g}]\,.\end{aligned} \tag{31}$$

Equation (30) then follows from Lemma 11. □

Remark 14. *When $\alpha > 1$ the biases of the two terms in equation (31) compete and so we can not obtain a bias bound via the above method.*

Similarly, we have:

Corollary 15 (f-Divergence Bias Bound)**.**

$$\mathbb{E}[E_{Q_n}[g] - \Lambda_f^{P_n}[g]] \geq E_Q[g] - \Lambda_f^P[g]$$

for all $g \in \mathcal{M}_b(\Omega)$ and

$$\mathbb{E}[D_f^\Gamma(Q_n\|P_n)] \geq D_f^\Gamma(Q\|P)\,. \tag{32}$$

E Experimental Details

Here, we describe the experimental setup and architectural details for the experiments presented in the paper. All experiments were performed using a fully-connected, feed-forward neural network of three hidden layers. We chose the number of units per layer to be 16, 16 and 8 for the one-dimensional Gaussians synthetic example as well as for the biological data example and $\tanh(\cdot)$ as activation function for hidden and output layers. In addition, we use gradient penalty with $\lambda_{\text{gp}} = 0.1$ during training using the biological data sets, for improved stability. We apply Adam optimizer as the training algorithm, with learning rate $\lambda_{\text{lr}} = 0.0005$ for the one-dimensional experiments, while the number of iterations is $N_{\text{it}} = 20000$ and batch size is $m = 124$. Due to slower convergence, the respective values for the biological data set are $\lambda_{\text{lr}} = 0.005$ and $N_{\text{it}} = 200000$. Moreover, we set a large value to the batch size $m = 1024$, so that samples from the tails are included in the statistical

average with high probability at each step. For 40-dimensional Gaussians, we used three hidden layers with 32, 32 and 16 hidden units respectively for DNE methods. The learning rate is set to 0.0001. For other methods, we chose the parameters as provided in the repository [1]. We follow the same architectures for the application of speech synthesis as mentioned in (Paul et al., 2021). Implementation is carried out using TensorFlow and code will be made available upon acceptance.

F FURTHER RESULTS ON SYNTHETIC DATA

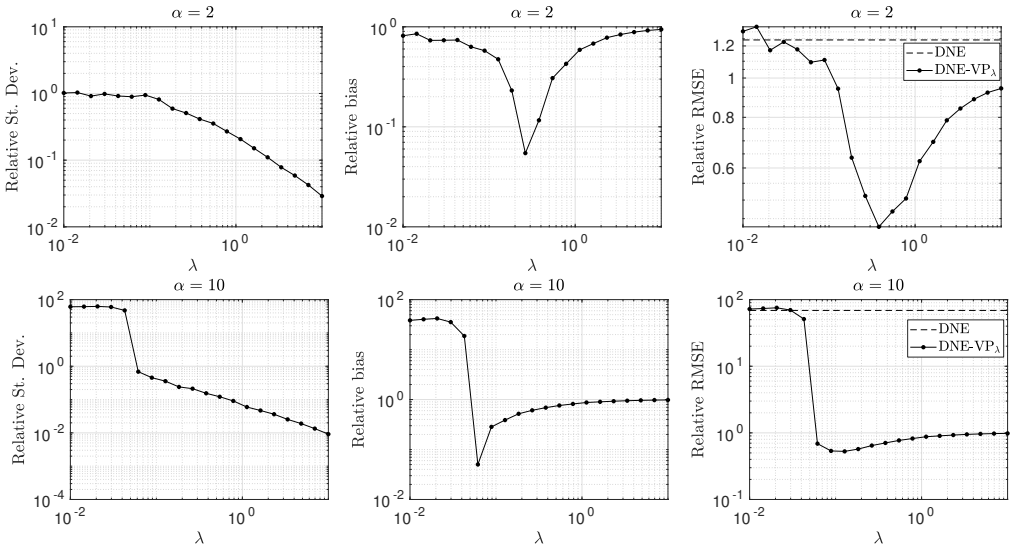

Figure 1: Relative standard deviation (left panels), relative bias (middle panels) and relative MSE (right panels) for the same experiment as in Figure 1 from the main text.

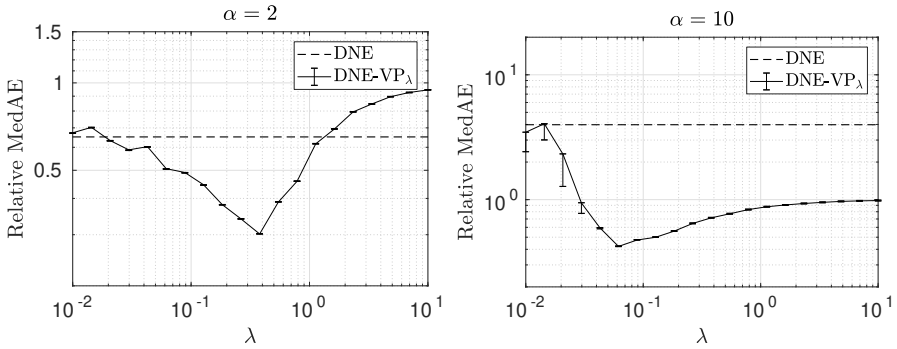

Figure 2: Same as Figure 1 from main text (middle columns) along with the 25% and 75% quartile interval.

REFERENCES

P. Billingsley. *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley, 2012. ISBN 9781118341919. URL https://books.google.com/books?id=a3gavZbxyJcC.

[1] https://github.com/Linear95/CLUB

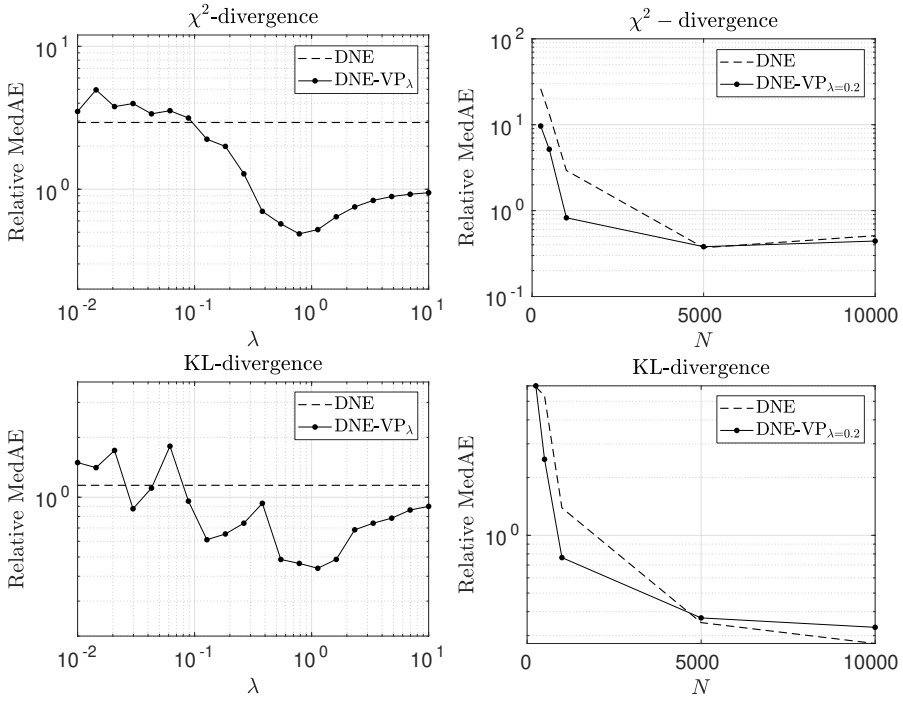

Figure 3: Comparison between the estimator without VP (DNE) and with VP (DNE-VP$_\lambda$) for f-divergence special cases of χ^2-divergence and KL-divergence, between two one-dimensional Gaussians with $Q = \mathcal{N}(0, 1.1)$ and $P = \mathcal{N}(0, 1)$. **Left column:** We compare the relative MedAE for varying penalty coefficient λ for sample size $N = 1000$, batch size 512, averaged over 50 i.i.d. runs. **Right column:** Relative MedAE for increasing sample size N similar to Figure 1.

P. Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Statistics. Wiley, 2013. ISBN 9781118625965. URL https://books.google.com/books?id=6ItqtwaWZZQC.

Jeremiah Birrell, Paul Dupuis, Markos A. Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet. (f, Γ)-Divergences: Interpolating between f-Divergences and Integral Probability Metrics. *arXiv e-prints*, art. arXiv:2011.05953, November 2020.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signal Systems*, 2:303–314, 1989.

G.B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2013. ISBN 9781118626399. URL https://books.google.com/books?id=wI4fAwAAQBAJ.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1QRgziT-.

Dipjyoti Paul, Sankar Mukherjee, Yannis Pantazis, and Yannis Stylianou. A Universal Multi-Speaker Multi-Style Text-to-Speech via Disentangled Representation Learning Based on Rényi Divergence Minimization. In *Proc. Interspeech 2021*, pp. 3625–3629, 2021. doi: 10.21437/Interspeech.2021-660.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R.G. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(70):2389–2410, 2011. URL http://jmlr.org/papers/v12/sriperumbudur11a.html.

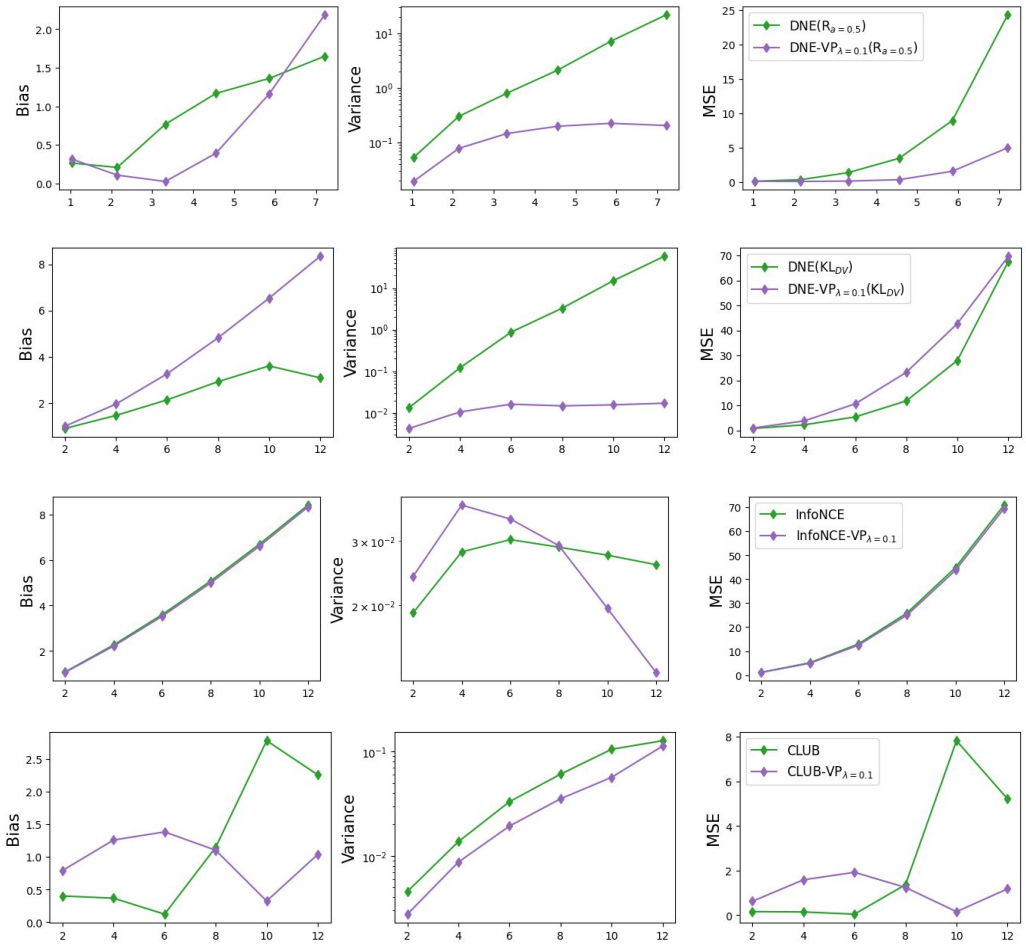

Figure 4: Estimation quality comparison of MI estimators considering bias, variance and mean square estimation error (MSE=bias^2+variance) for DNE, InfoNCE and CLUB variants. The horizontal axis shows the exact divergence values. The results are averaged over 20 i.i.d. runs. The results show improvements in the variance reduction front when VP is employed leading to lower MSE errors.

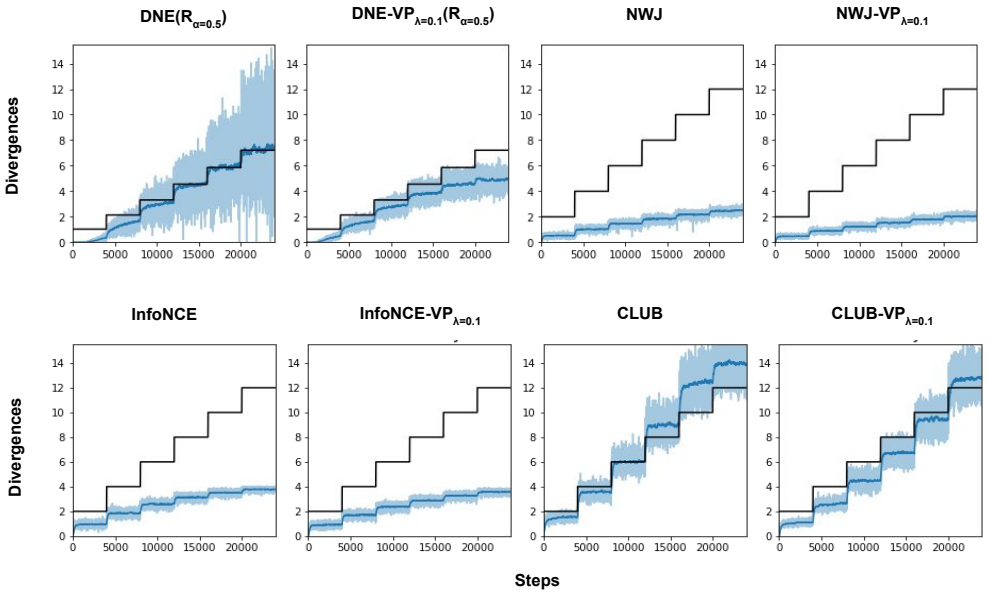

Figure 5: Performance comparison of several MI estimation approaches on a 40-dimensional correlated Gaussian random vector. Panels with $R_{\alpha=0.5}$ in their titles present the Rényi-based MI with $\alpha = 0.5$ whereas the rest of the methods estimate the standard MI (i.e., the KL divergence). Batch size is 64 and number of sample size set to 256K.

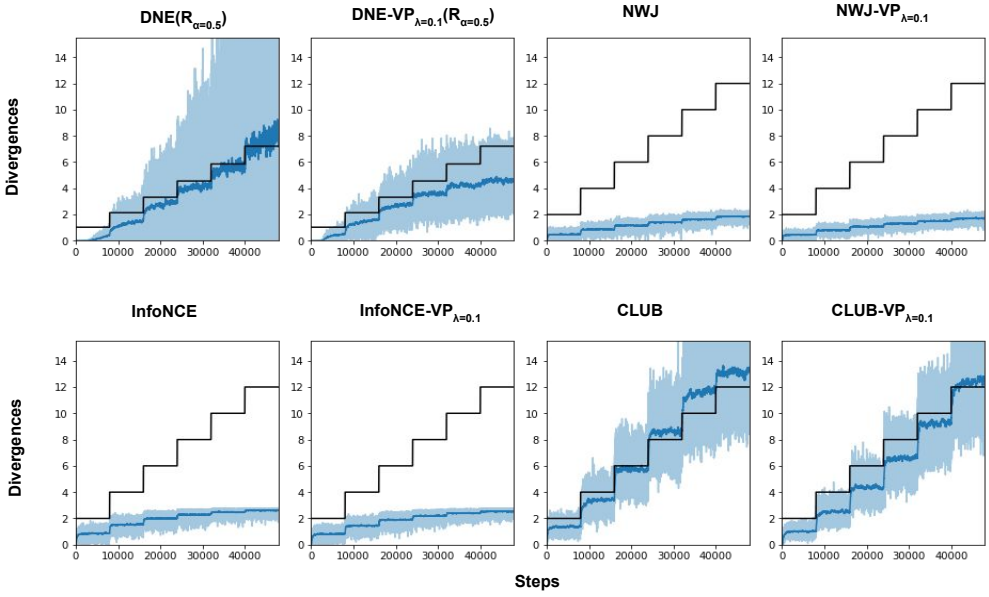

Figure 6: Performance comparison of several MI estimation approaches on a 40-dimensional correlated Gaussian random vector. Batch size is 16 and number of sample size set to 512K.

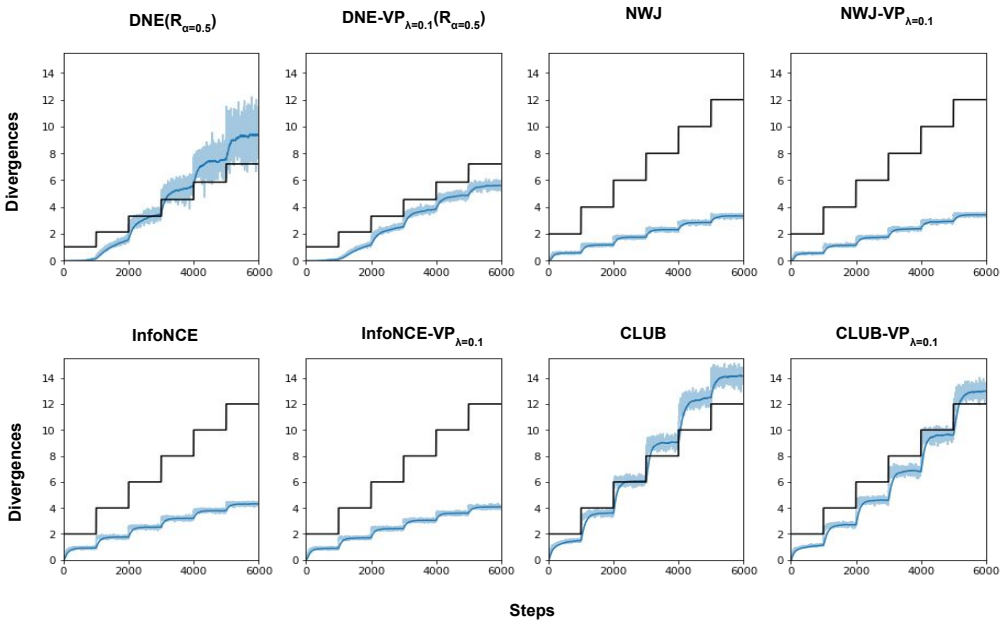

Figure 7: Performance comparison of several MI estimation approaches on a 40-dimensional correlated Gaussian random vector. Batch size is 512 and number of sample size set to 256K.

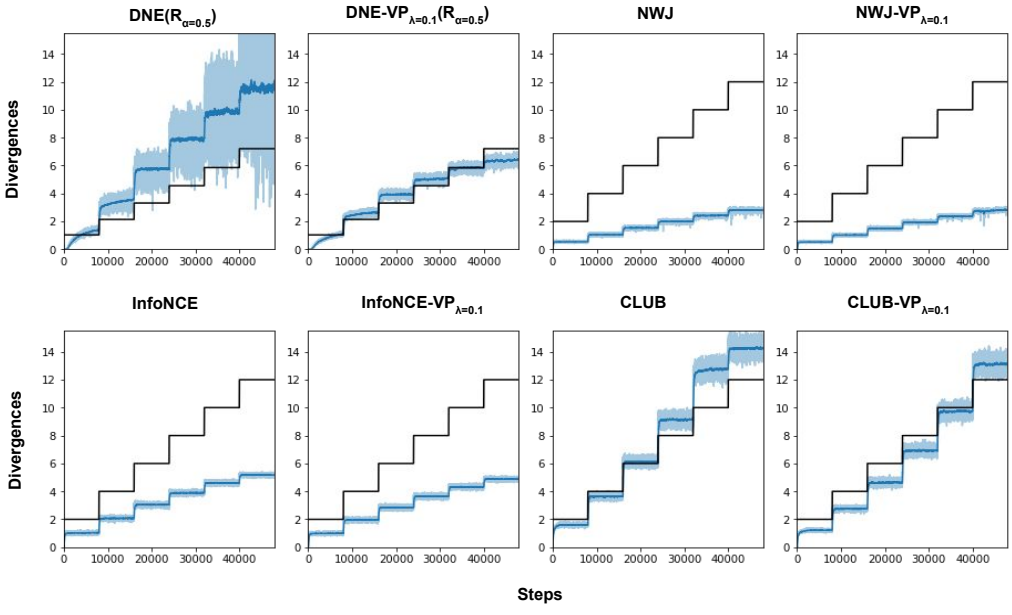

Figure 8: Performance comparison of several MI estimation approaches on a 40-dimensional correlated Gaussian random vector. Batch size is 512 and number of sample size set to 4M.

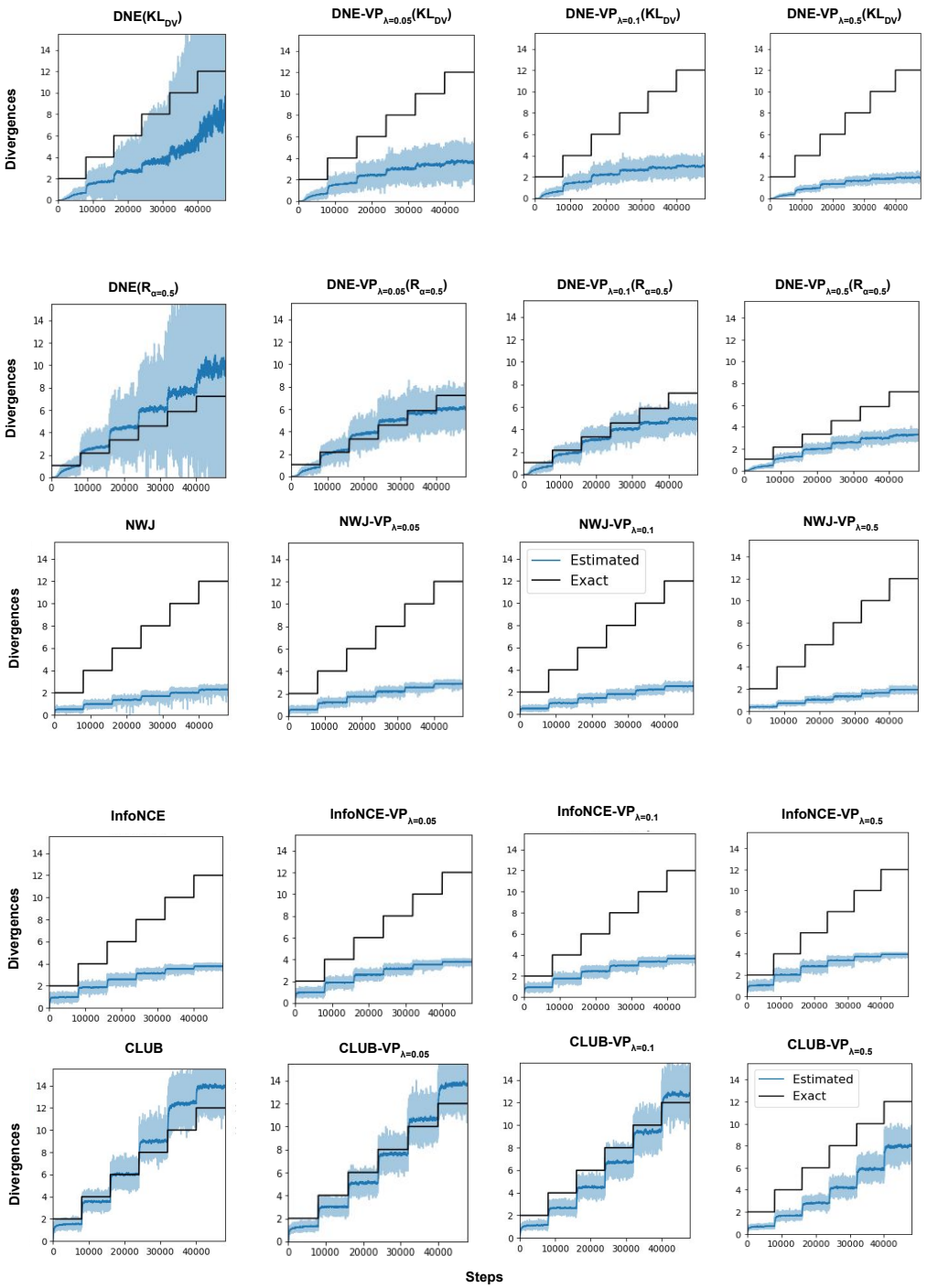

Figure 9: Performance comparison of several MI estimation approaches on a 40-dimensional correlated Gaussian random vector with different λ values. Batch size is 64 and number of sample size set to 512K.

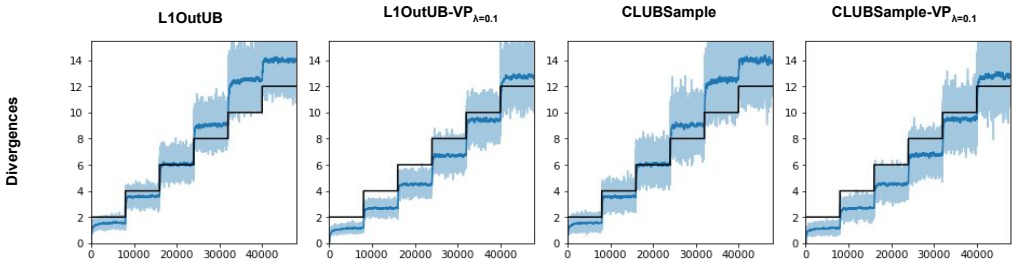

Figure 10: Performance comparison of several MI estimation approaches on a 40-dimensional correlated Gaussian random vector. Batch size is 64 and number of sample size set to 512K.