

## A Proofs of Propositions

For Lemmas 1 and 2, we first introduce some additional quantities to enable more compact notation; in particular, we adopt the value-function terminology from imitation learning. The following definitions are standard and immediate from the mapping given by Corollary 6, but are explicitly stated here for completeness. Recall that  $f_s : \mathcal{H} \times \mathcal{X} \rightarrow [-c, c]$  for some finite  $c$ . At any state  $h_t$ , define the “value function” to be the (forward-looking) expected sum of future quantities  $f_s(h_u, x_u)$  for  $u = t, \dots, T$ . Specifically, let  $V_{s,t}^{\pi_\theta}(h) : \mathcal{H} \rightarrow [-cT, cT]$  and  $Q_{s,t}^{\pi_\theta}(h, x) : \mathcal{H} \times \mathcal{X} \rightarrow [-cT, cT]$  be given as follows:

$$V_{s,t}^{\pi_\theta}(h) := \mathbb{E}_{\tau \sim p_\theta} [\sum_{u=t}^T f_s(h_u, x_u) | h_t = h] \quad (13)$$

$$Q_{s,t}^{\pi_\theta}(h, x) := \mathbb{E}_{\tau \sim p_\theta} [\sum_{u=t}^T f_s(h_u, x_u) | h_t = h, x_t = x] \quad (14)$$

where the notation for both  $V_{s,t}^{\pi_\theta}$  and  $Q_{s,t}^{\pi_\theta}$  is explicit as to their dependence on the policy  $\pi_\theta$  being followed, the source  $s$  under consideration, and the time  $t$ —unlike in typical imitation learning, we operate in a non-stationary (and non-Markovian) setting. For Lemma 1, we require an additional result:

**Lemma 7 (Expected Quality Difference)**  $\Delta \bar{F}_s(\theta) = T \mathbb{E}_{\substack{h \sim \mu_s \\ x \sim \pi_s(\cdot|h)}} Q_{s,t}^{\pi_\theta}(h, x) - T \mathbb{E}_{\substack{h \sim \mu_s \\ x \sim \pi_\theta(\cdot|h)}} Q_{s,t}^{\pi_\theta}(h, x)$ .

*Proof.* From Definition 2,

$$\Delta \bar{F}_s(\theta) = \mathbb{E}_{\tau \sim p_s} \sum_t f_s(h_t, x_t) - \mathbb{E}_{\tau \sim p_\theta} \sum_t f_s(h_t, x_t) \quad (15)$$

$$= \mathbb{E}_{\tau \sim p_s} \sum_t (f_s(h_t, x_t) + V_{s,t}^{\pi_\theta}(h_t) - V_{s,t}^{\pi_\theta}(h_t)) - \mathbb{E}_{\tau \sim p_\theta} \sum_t f_s(h_t, x_t) \quad (16)$$

$$= \mathbb{E}_{\tau \sim p_s} \sum_t (f_s(h_t, x_t) + V_{s,t+1}^{\pi_\theta}(h_{t+1}) - V_{s,t}^{\pi_\theta}(h_t)) \quad (17)$$

$$= \mathbb{E}_{\tau \sim p_s} \sum_t (Q_{s,t}^{\pi_\theta}(h_t, x_t) - V_{s,t}^{\pi_\theta}(h_t)) \quad (18)$$

$$= T \mathbb{E}_{h \sim \mu_s, x \sim \pi_s(\cdot|h)} (Q_{s,t}^{\pi_\theta}(h, x) - V_{s,t}^{\pi_\theta}(h)) \quad (19)$$

$$= T \mathbb{E}_{h \sim \mu_s, x \sim \pi_s(\cdot|h)} Q_{s,t}^{\pi_\theta}(h, x) - T \mathbb{E}_{h \sim \mu_s, x \sim \pi_\theta(\cdot|h)} Q_{s,t}^{\pi_\theta}(h, x) \quad (20)$$

where (16) to (17) telescopes terms, and we use the fact  $V_{s,T+1}^{\pi_\theta}(h) = 0$ . This derivation can be viewed as a non-stationary, non-Markovian analogue of the “performance difference” result in [103].  $\square$

**Lemma 1** Let  $\max_{f \in \mathbb{R}^{\mathcal{H} \times \mathcal{X}}} (\mathbb{E}_{\substack{h \sim \mu_s \\ x \sim \pi_s(\cdot|h)}} f(h, x) - \mathbb{E}_{\substack{h \sim \mu_s \\ x \sim \pi_\theta(\cdot|h)}} f(h, x)) \leq \epsilon$ . Then  $\Delta \bar{F}_s(\theta) \in O(T^2 \epsilon)$ .

*Proof.* From Lemma 7,

$$\Delta \bar{F}_s(\theta) = T \mathbb{E}_{h \sim \mu_s, x \sim \pi_s(\cdot|h)} Q_{s,t}^{\pi_\theta}(h, x) - T \mathbb{E}_{h \sim \mu_s, x \sim \pi_\theta(\cdot|h)} Q_{s,t}^{\pi_\theta}(h, x) \quad (21)$$

$$= T \mathbb{E}_{h \sim \mu_s} [\mathbb{E}_{x \sim \pi_s(\cdot|h)} Q_{s,t}^{\pi_\theta}(h, x) - \mathbb{E}_{x \sim \pi_\theta(\cdot|h)} Q_{s,t}^{\pi_\theta}(h, x)] \quad (22)$$

$$\leq \max_{Q \in [-cT, cT]^{\mathcal{H} \times \mathcal{X}}} T \mathbb{E}_{h \sim \mu_s} [\mathbb{E}_{x \sim \pi_s(\cdot|h)} Q(h, x) - \mathbb{E}_{x \sim \pi_\theta(\cdot|h)} Q(h, x)] \quad (23)$$

$$\leq \max_{f \in \mathbb{R}^{\mathcal{H} \times \mathcal{X}}} T \mathbb{E}_{h \sim \mu_s} [\mathbb{E}_{x \sim \pi_s(\cdot|h)} T f(h, x) - \mathbb{E}_{x \sim \pi_\theta(\cdot|h)} T f(h, x)] \quad (24)$$

$$\leq T^2 \epsilon \quad (25)$$

where the final inequality applies the assumption from the lemma. Note that this is similar in spirit to various results for error accumulation in imitation learning through behavioral cloning. The most well-known one is [61], where a quadratic bound is given with respect to the *probability* that the learned policy makes a small mistake. Another well-known one is in [104], where the bound is given with respect to *sample complexity*. Here, in order to motivate our perspective from the notion of expected quality difference, our bound is given with respect to the *moment-matching* discrepancy, and can be interpreted as a non-Markovian variant of the “off-policy upper bound” result in [105].  $\square$

**Lemma 2** Let  $\max_{f \in \mathbb{R}^{\mathcal{H} \times \mathcal{X}}} (\mathbb{E}_{\substack{h \sim \mu_s \\ x \sim \pi_s(\cdot|h)}} f(h, x) - \mathbb{E}_{\substack{h \sim \mu_\theta \\ x \sim \pi_\theta(\cdot|h)}} f(h, x)) \leq \epsilon$ . Then  $\Delta \bar{F}_s(\theta) \in O(T \epsilon)$ .

*Proof.* From Definition 1,

$$\Delta \bar{F}_s(\theta) = \mathbb{E}_{\tau \sim p_s} \sum_t f_s(h_t, x_t) - \mathbb{E}_{\tau \sim p_\theta} \sum_t f_s(h_t, x_t) \quad (26)$$

$$= T \mathbb{E}_{h \sim \mu_s, \pi_s(\cdot|h)} f_s(h, x) - T \mathbb{E}_{h \sim \mu_\theta, \pi_\theta(\cdot|h)} f_s(h, x) \quad (27)$$

$$\leq \max_{f \in [-c, c]^{\mathcal{H} \times \mathcal{X}}} (T \mathbb{E}_{h \sim \mu_s, x \sim \pi_s(\cdot|h)} f(h, x) - T \mathbb{E}_{h \sim \mu_\theta, x \sim \pi_\theta(\cdot|h)} f(h, x)) \quad (28)$$

$$\leq \max_{f \in \mathbb{R}^{\mathcal{H} \times \mathcal{X}}} (T \mathbb{E}_{h \sim \mu_s, x \sim \pi_s(\cdot|h)} f(h, x) - T \mathbb{E}_{h \sim \mu_\theta, x \sim \pi_\theta(\cdot|h)} f(h, x)) \quad (29)$$

$$\leq T \epsilon \quad (30)$$

where the final inequality applies the assumption from the lemma. Note that this is similar in spirit to various results for error accumulation in imitation learning through distribution matching. For instance, [106] shows a bound in terms of *divergences* in occupancy measures, while [104] shows a bound in terms of *sample complexity*. Here, in order to motivate our perspective from the notion of expected quality difference, our bound is given with respect to the *moment-matching* discrepancy, and can similarly be interpreted as a non-Markovian variant of the “reward upper bound” in [105].  $\square$

For Propositions 3 and 4, we use the fact that training the structured classifier (Definition 2) using the energy loss (Equation 11) amounts to a specific form of (sequence-wise) noise-contrastive estimation, and where the “noise”  $p_\theta$  employed happens to be adaptively trained via the policy loss (Equation 12):

**Proposition 3 (Global Optimality)** Let  $f_\phi \in \mathbb{R}^{\mathcal{H} \times \mathcal{X}}$ , and let  $p_\theta \in \Delta(\mathcal{T})$  be any distribution satisfying positivity:  $p_s(\tau) > 0 \Rightarrow p_\theta(\tau) > 0$  (this does not require  $\pi_\theta$  be optimal for  $f_\phi$ ). Then  $\mathcal{L}_{\text{energy}}(\phi; \theta)$  is globally minimized at  $F_\phi(\cdot) - \log Z_\phi = \log p_s(\cdot)$ , whence  $p_\phi$  is self-normalized with unit integral.

*Proof.* Briefly, a noise-contrastive estimator [79] operates as follows: Suppose we have some data  $y \in \mathcal{Y}$  distributed as  $p_{\text{data}}(y)$ . Consider that we wish to learn a model distribution  $p_{\text{model}}$ , as follows:

$$p_{\text{model}}(y; a, b) := \tilde{p}_{\text{model}}(y; a) \exp(b) \quad (31)$$

parameterized by  $a$  and  $b$ , where we emphasize that the model is not necessarily normalized as  $b$  is simply a learnable parameter. Also denote any noise distribution that can be sampled and evaluated:

$$p_{\text{noise}}(y; c) \quad (32)$$

parameterized by  $c$ . Now, define a classifier  $d(\cdot; a, b, c)$  as follows, which we shall train to discriminate between  $p_{\text{data}}$  and  $p_{\text{noise}}$ —that is, given some  $y$ , to represent the (posterior) probability that it is real:

$$d(y; a, b, c) := \sigma(\log p_{\text{model}}(y; a, b) - \log p_{\text{noise}}(y; c)) \quad (33)$$

where  $\sigma$  indicates the usual sigmoid function, i.e.  $\sigma(u) := 1/(1 + \exp(-u))$  for any  $u \in \mathbb{R}$ . The noise contrastive estimator maximizes the likelihood of the parameters  $a, b$  in  $d$  given  $p_{\text{data}}$  and  $p_{\text{noise}}$ :

$$\mathcal{L}_{\text{class}}(a, b; c) := -\mathbb{E}_{y \sim p_{\text{data}}} \log d(y; a, b, c) - \mathbb{E}_{y \sim p_{\text{noise}}} \log(1 - d(y; a, b, c)) \quad (34)$$

In this optimization problem, a basic result is that  $\mathcal{L}_{\text{class}}$  attains a minimum at  $\log p_{\text{model}} = \log p_{\text{data}}$  and that there are no other minima if  $p_{\text{noise}}$  is chosen such that  $p_{\text{data}}(y) > 0 \Rightarrow p_{\text{noise}}(y) > 0$  holds: see the “nonparametric estimation” result in [52]. Now, let us consider the following correspondence:

$$(\mathcal{Y}, p_{\text{data}}, \tilde{p}_{\text{model}}(\cdot; a), b, p_{\text{noise}}(\cdot; c)) := (\mathcal{T}, p_s, \tilde{p}_\phi, -\log Z_\phi, p_\theta) \quad (35)$$

In other words, let the underlying space be that of trajectories  $\mathcal{Y} := \mathcal{T}$ ; let the data distribution be  $p_{\text{data}} := p_s$ ; let the model distribution be given by the un-normalized energy model  $\tilde{p}_{\text{model}}(\cdot; a) := \tilde{p}_\phi$  and partition function  $b = -\log Z_\phi$ ; and let the noise distribution be given by rollouts of the policy,  $p_{\text{noise}}(\cdot; c) := p_\theta$ . Then it is easy to see that the classifier and its loss function correspond as follows:

$$\begin{aligned} d_{\theta, \phi}(\cdot) &= d(\cdot; a, b, c) \\ \mathcal{L}_{\text{energy}}(\phi; \theta) &= \mathcal{L}_{\text{class}}(a, b; c) \end{aligned} \quad (36)$$

But then the optimality result above directly maps to the statement that  $\mathcal{L}_{\text{energy}}$  is globally minimized at  $F_\phi(\cdot) - \log Z_\phi = \log p_s(\cdot)$  assuming that the positivity condition  $p_s(\tau) > 0 \Rightarrow p_\theta(\tau) > 0$  holds. Technicality: Note that here  $\tilde{p}_\phi$  is constrained as  $F_\phi(\tau) := \sum_t f_\phi(h_t, x_t)$  instead of being arbitrarily parameterizable, but this does not affect realizability as we assumed that  $F_s(\tau) := \sum_t f_s(h_t, x_t)$ .  $\square$

**Proposition 4 (Asymptotic Consistency)** Let  $\phi^*$  denote the minimizer for  $\mathcal{L}_{\text{energy}}(\phi; \theta)$ , and let  $\hat{\phi}_M^*$  denote the minimizer for its finite-data approximation—that is, where the expectations over  $p_s$  and  $p_\theta$  are approximated by  $M$  samples. Then under some mild conditions, as  $M$  increases  $\hat{\phi}_M^* \xrightarrow{p} \phi^*$ .

*Proof.* Continuing the exposition above, let  $\mathcal{L}_{\text{class}}^M(a, b; c)$  indicate the finite-data approximation of  $\mathcal{L}_{\text{class}}(a, b; c)$ —that is, by using  $M$  samples to approximate the true expectations over  $p_{\text{data}}$  and  $p_{\text{noise}}$ :

$$\mathcal{L}_{\text{class}}^M(a, b; c) := -\frac{1}{M} \sum_{m=1}^M \log d(y_{\text{data}}^{(m)}; a, b, c) - \frac{1}{M} \sum_{m=1}^M \log(1 - d(y_{\text{noise}}^{(m)}; a, b, c)) \quad (37)$$

where the samples are drawn as  $y_{\text{data}}^{(m)} \sim p_{\text{data}}$  and  $y_{\text{noise}}^{(m)} \sim p_{\text{noise}}$ . Consider the following conditions:

1. Positivity:  $p_{\text{data}}(y) > 0 \Rightarrow p_{\text{noise}}(y) > 0$ ;
2. Uniform convergence:  $\sup_{a,b} |\mathcal{L}_{\text{class}}^M(a, b; c) - \mathcal{L}_{\text{class}}(a, b; c)| \xrightarrow{p} 0$ ; and
3. The following matrix is full-rank:  $\mathcal{I} := \int g(y)g(y)^\top p_{\text{data}}(y)p_{\text{noise}}(y)/(p_{\text{data}}(y) + p_{\text{noise}}(y))dy$ , where  $g(y) := \nabla_{(a,b)} \log p_{\text{model}}(y; a, b)|_{(a^*, b^*)}$  and  $a^*, b^*$  denote the optimal values of the model.

Note that (1) is same as before, and (2) and (3) are analogous to standard assumptions in maximum likelihood estimation. Let  $\hat{a}_M^*, \hat{b}_M^*$  denote the minimizers for  $\mathcal{L}_{\text{class}}^M(a, b; c)$ . Under the preceding conditions, another basic result is that  $(\hat{a}_M^*, \hat{b}_M^*)$  converges in probability to  $(a^*, b^*)$  as  $M$  grows: see the ‘‘consistency’’ result in [52]. But continuing the correspondence from before, it is easy to see that

$$\mathcal{L}_{\text{energy}}^M(\phi; \theta) = \mathcal{L}_{\text{class}}^M(a, b; c) \quad (38)$$

where we similarly define  $\mathcal{L}_{\text{energy}}^M(\phi; \theta)$  to be the finite-data approximation of  $\mathcal{L}_{\text{energy}}(\phi; \theta)$ —that is, by using  $M$  samples to approximate the true expectations over  $p_s$  and  $p_\theta$ , and  $y_s^{(m)} \sim p_s$  and  $y_\theta^{(m)} \sim p_\theta$ :

$$\mathcal{L}_{\text{energy}}^M(\phi; \theta) := -\frac{1}{M} \sum_{m=1}^M \log d_{\theta, \phi}(\tau_s^{(m)}) - \frac{1}{M} \sum_{m=1}^M \log(1 - d_{\theta, \phi}(\tau_\theta^{(m)})) \quad (39)$$

which directly maps the above convergence result to the statement that as  $M$  increases  $\hat{\phi}_M^* \xrightarrow{p} \phi^*$ .  $\square$

**Proposition 5 (Gradient Equality)** Let  $\phi_k$  be the value taken by  $\phi$  after the  $k$ -th gradient update, and let  $\theta_k^*$  denote the associated minimizer for  $\mathcal{L}_{\text{policy}}(\theta; \phi_k)$ . Suppose  $p_\phi$  is already normalized; then

$$\nabla_\phi \mathcal{L}_{\text{energy}}(\phi; \theta_k^*) = -\frac{T}{2} \nabla_\phi \mathcal{L}(\theta_k^*, \phi)$$

That is, at  $\theta_k^*$  the energy gradient (of Equation 11) recovers the original gradient (from Equation 7). In the general case, suppose  $p_\phi$  is un-normalized, such that  $p_{\theta_k^*} = p_\phi / K_\phi$  for some constant  $K_\phi$ ; then

$$\nabla_\phi \mathcal{L}_{\text{energy}}(\phi; \theta_k^*) = \frac{TK_\phi}{K_\phi + 1} \mathbb{E}_{\substack{h \sim \mu_{\theta_k^*} \\ x \sim \pi_{\theta_k^*}(\cdot|h)}} \nabla_\phi f_\phi(h, x) - \frac{T}{K_\phi + 1} \mathbb{E}_{\substack{h \sim \mu_s \\ x \sim \pi_s(\cdot|h)}} \nabla_\phi f_\phi(h, x)$$

*Proof.* From Equation 11,

$$\nabla_\phi \mathcal{L}_{\text{energy}}(\phi; \theta_k^*) = \nabla_\phi \left( -\mathbb{E}_{\tau \sim p_s} \log d_{\theta_k^*, \phi}(\tau) - \mathbb{E}_{\tau \sim p_{\theta_k^*}} \log(1 - d_{\theta_k^*, \phi}(\tau)) \right) \quad (40)$$

$$= \nabla_\phi \left( -\mathbb{E}_{\tau \sim p_s} \log \frac{p_\phi(\tau)}{p_\phi(\tau) + p_{\theta_k^*}(\tau)} - \mathbb{E}_{\tau \sim p_{\theta_k^*}} \log \frac{p_{\theta_k^*}(\tau)}{p_\phi(\tau) + p_{\theta_k^*}(\tau)} \right) \quad (41)$$

$$= -\mathbb{E}_{\tau \sim p_s} \nabla_\phi (\log p_\phi(\tau) - \log(p_\phi(\tau) + p_{\theta_k^*}(\tau))) + \mathbb{E}_{\tau \sim p_{\theta_k^*}} \nabla_\phi \log(p_\phi(\tau) + p_{\theta_k^*}(\tau)) \quad (42)$$

$$= -\mathbb{E}_{\tau \sim p_s} \left[ \nabla_\phi \log p_\phi(\tau) - \frac{\nabla_\phi p_\phi(\tau)}{p_\phi(\tau) + p_{\theta_k^*}(\tau)} \right] + \mathbb{E}_{\tau \sim p_{\theta_k^*}} \frac{\nabla_\phi p_\phi(\tau)}{p_\phi(\tau) + p_{\theta_k^*}(\tau)} \quad (43)$$

$$= -\mathbb{E}_{\tau \sim p_s} \left[ \nabla_\phi \log p_\phi(\tau) - \frac{p_\phi(\tau) \nabla_\phi \log p_\phi(\tau)}{p_\phi(\tau) + p_{\theta_k^*}(\tau)} \right] + \mathbb{E}_{\tau \sim p_{\theta_k^*}} \frac{p_\phi(\tau) \nabla_\phi \log p_\phi(\tau)}{p_\phi(\tau) + p_{\theta_k^*}(\tau)} \quad (44)$$

$$= -\mathbb{E}_{\tau \sim p_s} \left[ \nabla_\phi \log p_\phi(\tau) - \frac{1}{2} \nabla_\phi \log p_\phi(\tau) \right] + \frac{1}{2} \mathbb{E}_{\tau \sim p_{\theta_k^*}} \nabla_\phi \log p_\phi(\tau) \quad (45)$$

$$= \frac{1}{2} \mathbb{E}_{\tau \sim p_{\theta_k^*}} \nabla_\phi \log p_\phi(\tau) - \frac{1}{2} \mathbb{E}_{\tau \sim p_s} \nabla_\phi \log p_\phi(\tau) \quad (46)$$

$$= \frac{1}{2} \mathbb{E}_{\tau \sim p_{\theta_k^*}} \nabla_\phi (\log \tilde{p}_\phi(\tau) - \log Z_\phi) - \frac{1}{2} \mathbb{E}_{\tau \sim p_s} \nabla_\phi (\log \tilde{p}_\phi(\tau) - \log Z_\phi) \quad (47)$$

$$= \frac{1}{2} (\mathbb{E}_{\tau \sim p_{\theta_k^*}} \nabla_\phi \sum_t f_\phi(h_t, x_t) - \mathbb{E}_{\tau \sim p_s} \nabla_\phi \sum_t f_\phi(h_t, x_t)) \quad (48)$$

$$= \frac{1}{2} (T \mathbb{E}_{h \sim \mu_{\theta_k^*}, x \sim \pi_{\theta_k^*}(\cdot|h)} \nabla_\phi f_\phi(h, x) - T \mathbb{E}_{h \sim \mu_s, x \sim \pi_s(\cdot|h)} \nabla_\phi f_\phi(h, x)) \quad (49)$$

$$= -\frac{T}{2} \nabla_\phi \mathcal{L}(\theta_k^*, \phi) \quad (50)$$

where the fourth and fifth lines repeatedly use the identity  $\nabla_z p_z \equiv p_z \nabla_z \log p_z$  for any  $p_z$  parameterized by  $z$ , and the sixth line uses the fact that the current value of  $\theta$  (i.e.  $\theta_k^*$ ) is the minimizer for  $\mathcal{L}_{\text{policy}}(\theta; \phi)$  at the current value of  $\phi$  (i.e.  $\phi_k$ ), hence it must be the case that  $p_\theta = p_\phi$  at those values. Note that this assumes that  $p_\phi$  is normalized; in practice this will be approximately true, for instance if we pre-train  $\phi$  beforehand, using a fixed  $\pi_\theta$  pre-trained by maximum likelihood (see Appendix B).

In the more general case of any arbitrary un-normalized  $p_\phi$ , we only know  $p_{\theta_k^*} = \frac{1}{K_\phi} p_\phi$  for some constant  $K_\phi$ ; then we recover a generalized “weighted” version of Equation 7. From the fifth line above,

$$\nabla_\phi \mathcal{L}_{\text{energy}}(\phi; \theta_k^*) = -\mathbb{E}_{\tau \sim p_s} \left[ \nabla_\phi \log p_\phi(\tau) - \frac{p_\phi(\tau) \nabla_\phi \log p_\phi(\tau)}{p_\phi(\tau) + p_{\theta_k^*}(\tau)} \right] + \mathbb{E}_{\tau \sim p_{\theta_k^*}} \frac{p_\phi(\tau) \nabla_\phi \log p_\phi(\tau)}{p_\phi(\tau) + p_{\theta_k^*}(\tau)} \quad (51)$$

$$= -\mathbb{E}_{\tau \sim p_s} \left[ \nabla_\phi \log p_\phi(\tau) - \frac{K_\phi}{K_\phi + 1} \nabla_\phi \log p_\phi(\tau) \right] + \frac{K_\phi}{K_\phi + 1} \mathbb{E}_{\tau \sim p_{\theta_k^*}} \nabla_\phi \log p_\phi(\tau) \quad (52)$$

$$= \frac{K_\phi}{K_\phi + 1} \mathbb{E}_{\tau \sim p_{\theta_k^*}} \nabla_\phi \log p_\phi(\tau) - \frac{1}{K_\phi + 1} \mathbb{E}_{\tau \sim p_s} \nabla_\phi \log p_\phi(\tau) \quad (53)$$

$$= \frac{K_\phi}{K_\phi + 1} \mathbb{E}_{\tau \sim p_{\theta_k^*}} \nabla_\phi (\log \tilde{p}_\phi(\tau) - \log Z_\phi) - \frac{1}{K_\phi + 1} \mathbb{E}_{\tau \sim p_s} \nabla_\phi (\log \tilde{p}_\phi(\tau) - \log Z_\phi) \quad (54)$$

$$= \frac{K_\phi}{K_\phi + 1} \mathbb{E}_{\tau \sim p_{\theta_k^*}} \nabla_\phi \sum_t f_\phi(h_t, x_t) - \frac{1}{K_\phi + 1} \mathbb{E}_{\tau \sim p_s} \nabla_\phi \sum_t f_\phi(h_t, x_t) \quad (55)$$

$$= \frac{TK_\phi}{K_\phi + 1} \mathbb{E}_{h \sim \mu_{\theta_k^*}, x \sim \pi_{\theta_k^*}(\cdot|h)} \nabla_\phi f_\phi(h, x) - \frac{T}{K_\phi + 1} \mathbb{E}_{h \sim \mu_s, x \sim \pi_s(\cdot|h)} \nabla_\phi f_\phi(h, x) \quad (56)$$

This “weighting” is intuitive: If  $p_\phi$  is un-normalized such that  $K_\phi > 1$ , the energy loss automatically places higher weights on negative samples  $h \sim \mu_{\theta_k^*}, x \sim \pi_{\theta_k^*}(\cdot|h)$  to bring it down; conversely, if  $p_\phi$  is un-normalized such that  $K_\phi < 1$ , the energy loss places higher weights on positive samples  $h \sim \mu_s, x \sim \pi_s(\cdot|h)$  to bring it up. (If  $p_\phi$  is normalized, then  $K_\phi = 1$  and the weights are equal).  $\square$

**Note on Duality** Lemmas 1 and 2 are for building intuition, and we are *not* implying that Equations 3 and 5 are direct generalizations of the duality between maximum likelihood and maximum entropy. To be clear, we shall explain Equation 5; Equation 3 is similar. Consider first the case of finite  $\mathcal{X}$  (therefore finite  $\mathcal{T}$ ). In the usual linear case, given basis functions  $T(\tau)$ , the optimization problem is:

$$\arg \min_\theta E_{\tau \sim p_\theta} \log p_\theta(\tau) \quad \text{s.t.} \quad E_{\tau \sim p_s} T(\tau) = E_{\tau \sim p_\theta} T(\tau) \quad (57)$$

Internalizing the constraint, we may write  $\arg \min_\theta (E_{\tau \sim p_\theta} \log p_\theta(\tau) + \max_F (\langle F, E_{\tau \sim p_s} T(\tau) \rangle - \langle F, E_{\tau \sim p_\theta} T(\tau) \rangle))$ . To generalize to the nonlinear case, let us now specifically define the feature vector  $T(\tau)$  to be the indicator function (i.e. a finite-length vector, each zero-one entry of which corresponds to each element in  $\mathcal{T}$ ). Then note that  $\langle F, E_{\tau \sim p} T(\tau) \rangle = \langle F, p \rangle = E_{\tau \sim p} F(\tau)$ , therefore:

$$\arg \min_\theta (E_{\tau \sim p_\theta} \log p_\theta(\tau) + \max_F (E_{\tau \sim p_s} F(\tau) - E_{\tau \sim p_\theta} F(\tau))) \quad (58)$$

Finally, to arrive at Equation 5 we use the fact that  $E_{\tau \sim p} \sum_t f(h_t, x_t) = T E_{h \sim \mu, x \sim \pi(\cdot|h)} f(h, x)$ , as already noted. All in all, this is a “generalization” from the case of linearity in *known* basis functions, to the case of *unknown* basis functions—where we “linearize” the expression using indicator functions. But we cannot use the same logic to claim the same for *infinite*  $\mathcal{X}$  (which is the setting we operate in).

## B Details on Algorithm

**Policy Optimization** Recall the policy update (Equation 12); this corresponds to entropy-regularized reinforcement learning using  $f_\phi(h, x)$  as transition-wise reward function. Here we give a brief review of entropy-regularized reinforcement learning [45–47] in our context, as well as the practical method we employ (i.e. soft actor-critic). First, we introduce some standard notation. At any state  $h$ , define the (soft) “value function” to be the (forward-looking) expected sum of future rewards  $f_\phi(h, x)$  as well as entropies  $H(\pi(\cdot|h))$ . Specifically, let  $V_\phi^{\pi_\theta}(h)$  and  $Q_\phi^{\pi_\theta}(h, x)$  be given as follows (we omit explicit notation for  $t$ , as any influence of time is implicit through dependence on variable-length histories):

$$V_\phi^{\pi_\theta}(h) := \mathbb{E}_{\tau \sim p_\theta} [\sum_{u=t}^T f_\phi(h_u, x_u) + H(\pi_\theta(\cdot|h_u)) | h_t = h] \quad (59)$$

$$Q_\phi^{\pi_\theta}(h, x) := f_\phi(h, x) + \mathbb{E}_{\tau \sim p_\theta} [\sum_{u=t+1}^T f_\phi(h_u, x_u) + H(\pi_\theta(\cdot|h_u)) | h_t = h, x_t = x] \quad (60)$$

Let  $\pi_{\theta^*}$  denote the optimal policy (i.e. that minimizes loss  $\mathcal{L}_{\text{policy}}$ ), and  $Q_\phi^{\pi_{\theta^*}}$  its corresponding value function. An elementary result is that the optimal policy assigns probabilities to  $x$  proportional to the exponentiated expected returns of energy and entropy terms of all trajectories that begin with  $(h, x)$ :

$$\pi_{\theta^*}(x|h) = \frac{\exp(Q_\phi^{\pi_{\theta^*}}(h, x))}{\int_{\mathcal{X}} \exp(Q_\phi^{\pi_{\theta^*}}(h, x)) dx} \quad (61)$$

Now, for any transition policy  $\pi_\theta$  (i.e. not necessarily optimal with respect to  $f_\phi$ ), the value function  $Q_\phi^{\pi_\theta}$  is the unique fixed point of the following (soft) Bellman backup operator  $\mathbb{B}_\phi^{\pi_\theta} : \mathbb{R}^{\mathcal{H} \times \mathcal{X}} \rightarrow \mathbb{R}^{\mathcal{H} \times \mathcal{X}}$ :

$$(\mathbb{B}_\phi^{\pi_\theta} Q)(h, x) := f_\phi(h, x) + \mathbb{E}_{x' \sim \pi_\theta(\cdot|h')} [Q(h', x') - \log \pi_\theta(x'|h')] \quad (62)$$

and hence—in theory— $Q_\phi^{\pi_\theta}$  may be computed iteratively by repeatedly applying the operator  $\mathbb{B}_\phi^{\pi_\theta}$  starting from any function  $Q \in \mathbb{R}^{\mathcal{H} \times \mathcal{X}}$ ; this is referred to as the (soft) “policy evaluation” procedure.

Using  $Q_\phi^{\pi_\theta}$ , we may then perform (soft) “policy improvement” to update the policy  $\pi_\theta$  towards the exponential of its value function, and is guaranteed to result in an improved policy (in terms of  $Q_\phi^{\pi_\theta}$ ):

$$\theta' \leftarrow \arg \min_\theta D_{\text{KL}} \left( \pi_{\theta'}(\cdot|h) \parallel \frac{\exp(Q_\phi^{\pi_\theta}(h, \cdot))}{\int_{\mathcal{X}} \exp(Q_\phi^{\pi_\theta}(h, x)) dx} \right) \quad (63)$$

for all  $h \in \mathcal{H}$ . In theory, then, finding the optimal policy can be approached by repeatedly applying the above policy evaluation and policy improvement steps starting from any initial policy  $\pi_\theta$ ; this is referred to as (soft) “policy iteration”. However, in large continuous domains (such as  $\mathcal{H} \times \mathcal{X}$ ) doing this exactly is impossible, so we need to rely on function approximation for representing value functions.

**Practical Algorithm** Precisely, the soft actor-critic approach is to introduce a function approximator to represent the value function (i.e. the “critic”) parameterized by  $\psi$ , in addition to the policy itself (i.e. the “actor”) parameterized by  $\theta$ , and to alternate between optimizing both with stochastic gradient descent [57]. Specifically, the actor performs soft policy improvement steps as before, but now using  $Q_\psi$ :

$$\mathcal{L}_{\text{actor}}(\theta; \phi, \psi) := \mathbb{E}_{h \sim \mathcal{B}} \mathbb{E}_{x \sim \pi_\theta(\cdot|h)} [\log \pi_\theta(x|h) - Q_\psi(h, x)] \quad (64)$$

where  $\mathcal{B}$  is a replay buffer of samples generated by  $\pi_\theta$  (that is, instead of enumerating all  $h \in \mathcal{H}$ , we are relying on  $h \sim \mathcal{B}$ ). Note that the normalizing constant is dropped as it does not contribute to the gradient. The critic is trained to represent the value function by minimizing squared residual errors:

$$\mathcal{L}_{\text{critic}}(\psi; \phi) := \mathbb{E}_{h, x \sim \mathcal{B}} (Q_\psi(h, x) - Q_\psi^{\text{target}}(h, x))^2 \quad (65)$$

with (bootstrapped) targets:

$$Q_\psi^{\text{target}}(h, x) := f_\phi(h, x) + \mathbb{E}_{x' \sim \pi_\theta(\cdot|h')} [Q_\psi(h', x') - \log \pi_\theta(x'|h')] \quad (66)$$

Together, this provides a way to minimize the policy loss  $\mathcal{L}_{\text{policy}}$  (Equation 12). The complete Time-GCI algorithm simply alternates between this and minimizing the energy loss  $\mathcal{L}_{\text{energy}}$  (Equation 11):

$$\mathcal{L}_{\text{energy}}(\phi; \theta) := -\mathbb{E}_{\tau \sim p_s} \log d_{\theta, \phi}(\tau) - \mathbb{E}_{\tau \sim p_\theta} \log (1 - d_{\theta, \phi}(\tau)) \quad (67)$$

Hence in Algorithm 1, gradient updates for the energy, policy, and critic are interleaved with policy rollouts. Note that several standard approximations are being used. First, the replay buffer provides samples  $h \sim \mathcal{B}$  for optimizing the policy (in both actor and critic updates), instead of covering the entire space  $\mathcal{H}$  (which is uncountable). Second, in the energy loss negative samples  $\tau \sim \mathcal{B}$  are used in lieu of sampling fresh from  $p_\theta$  at every iteration (this is known to give the benefit of providing more diverse negative samples). Finally, also per usual samples from the dataset  $\tau \sim \mathcal{D}$  is used in lieu of  $p_s$ .

**Practical Considerations** First, in practice we must use a *vector representation* of histories  $h \in \mathcal{H}$ ; here we use RNNs to encode histories into fixed-length vectors, which can then be treated as regular “states” in continuous space. Like our choice of policy optimization, this is also an arbitrary design choice—we could just as conceivably have used e.g. temporal convolutions, attention mechanisms, etc.

Second, *interleaving* multiple gradient updates of different networks requires some care: In soft actor-critic itself, policy updates have to be sufficiently small, and/or critic updates have to be sufficiently frequent, to prevent divergence. The situation is analogous when interleaving this with energy gradient updates as well: Both actor and energy updates have to be sufficiently small, and/or critic updates have to be sufficiently frequent. That said, the energy updates are indeed decoupled from the policy updates: Regardless of how quickly/slowly the policy is learning, the energy can learn on their negative samples. In practice, we perform multiple critic updates for every update of the policy and energy functions.

Finally, note that in large continuous domains such as  $\mathcal{H} \times \mathcal{X}$  it is necessary to *pre-train* the networks beforehand such that optimization of the complete algorithm actually converges: On the one hand, the policy side requires a sufficiently good energy signal to actually make progress, and on the other hand, the energy side requires a sufficiently good policy providing challenging enough negative samples to actually make progress. Pre-training networks separately is standard in actor-critic methods (see for instance [35]); here we take a similar approach but with the addition of the energy update step as well:

1. Policy-only:  $\pi_\theta$  is pre-trained using maximum likelihood;
2. Energy-only:  $f_\phi$  is pre-trained using  $\mathcal{L}_{\text{energy}}(\phi; \theta)$ , holding  $\pi_\theta$  fixed;
3. Critic-only:  $Q_\psi$  is pre-trained using  $\mathcal{L}_{\text{critic}}(\psi; \phi)$ , holding  $\pi_\theta, f_\phi$  fixed; and finally,
4. All:  $f_\phi, \pi_\theta$ , and  $Q_\psi$  are trained on  $\mathcal{L}_{\text{energy}}(\phi; \theta), \mathcal{L}_{\text{actor}}(\theta; \phi, \psi)$ , and  $\mathcal{L}_{\text{critic}}(\psi; \phi)$  (cf. Algorithm 1).

## C Details on Experiments

**Benchmark Algorithms** Except where components are standardized (see below), we use the publicly available source code when constructing the benchmark algorithms; references are in the following:

- T-Forcing [5]: (straightforward MLE with ground-truth conditioning)
- P-Forcing [10]: [https://github.com/anirudh9119/LM\\_GANS](https://github.com/anirudh9119/LM_GANS)
- C-RNN-GAN [18]: <https://github.com/olofmogren/c-rnn-gan>
- COT-GAN [20]: <https://github.com/tianlinxu312/cot-gan>
- RC-GAN [21]: <https://github.com/ratschlab/RCGAN>
- TimeGAN [12]: <https://github.com/jsyoon0823/TimeGAN>

**Dataset Sources** We use the original source code for preprocessing sines and UCI datasets from TimeGAN (<https://github.com/jsyoon0823/TimeGAN>). For MIMIC-III, we extract 52 clinical covariates including vital signs (e.g. respiratory rate, heart rate, O2 saturation) and lab tests (e.g. glucose, hemoglobin, white blood cell count) aggregated every hour during their ICU stay up to 24 hours.

- Sines [5]: <https://github.com/jsyoon0823/TimeGAN>
- Energy [10]: [archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction](https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction)
- Gas [18]: [archive.ics.uci.edu/ml/datasets/Gas+sensor+array+temperature+modulation](https://archive.ics.uci.edu/ml/datasets/Gas+sensor+array+temperature+modulation)
- Metro [20]: [archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume](https://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume)
- MIMIC-III [21]: <https://physionet.org/content/mimiciii/1.4/>

**Encoder Networks** For fair comparison, analogous network components across all benchmarks share the same architecture where possible. In particular, all components taking  $h_t$  as input require an encoder network to construct fixed-length vector representations of variable-length histories  $(x_1, \dots, x_t)$ . To do so, these components use an LSTM network with a hidden layer of size 32 to compute hidden states for representing  $h$ . These are not shared: separate components have their own encoder networks.

**Task-Specific Networks** Then, for mapping from  $h_t$  and/or  $x_t$  to task-specific output variables, we use a fully-connected network with two hidden layers of size 32 and ELU activations. Where both  $h_t$  and  $x_t$  serve as inputs, their vectors are concatenated. For instance, the energy network for TimeGCI contains one such network for computing  $f_\phi$  (as well as trainable parameter for  $Z_\phi$ ). The same applies to the mapping from  $h_t$  and/or  $x_t$  to implicit generator outputs (as in C-RNN-GAN and RC-GAN), black-box discriminator output (as in P-Forcing, C-RNN-GAN, and RC-GAN), transition policies (as in T-Forcing and P-Forcing), and critic values (for TimeGCI). Note that TimeGAN and COT-GAN were designed with additional unique components/losses that operate as a unit (e.g. the embedding/recovery networks for generating/discriminating within latent space); we use their original source code.

**Replay Buffer** The replay buffer  $\mathcal{B}$  has a fixed size; once filled, new samples stored replace the oldest still in the buffer. Sampling from the buffer operates as follows: In updating the energy, we require  $\tau \sim \mathcal{B}$  (that is, in lieu of  $p_\theta$ , cf. Equation 67); this is done by randomly sampling a batch of trajectories from the replay buffer, without replacement. In policy the actor, we require  $h \sim \mathcal{B}$  (cf. Equation 64); this is done by first randomly sampling a batch of trajectories from the replay buffer, and then randomly sampling a cutoff time  $t$  to obtain a batch of subsequences  $h_t$ . Finally, in updating the critic, we require  $h, x \sim \mathcal{B}$  (cf. Equation 65); this is similarly done by first randomly sampling a batch of trajectories from the replay buffer, then randomly sampling a cutoff  $t$  to yield a batch of  $(h_t, x_t)$  pairs.

**Hyperparameters** In all experiments, we use the following hyperparameters for TimeGCI and for benchmarks (wherever applicable): The replay buffer is of size  $|\mathcal{B}| = 10,000$ . The hidden dimension of all encoder and task-specific networks is 32. The entropy regularization (in the actor/policy loss) is  $\alpha = 0.2$ . The policy network is pre-trained for 2,000 steps, energy for 4,000, and critic for 20,000. The complete algorithm is trained for up to 50,000 steps with checkpointing and early stopping (triggerable every 1,000 iterations, if performance does not improve). The Adam optimizer is used for all losses, and batch size  $M = 64$ . Learning rates: For energy networks  $\lambda_{\text{energy}} = 0.0001$ , for policy networks  $\lambda_{\text{policy}} = 0.0001$  (same for implicit generator networks), for critic networks  $\lambda_{\text{critic}} = 0.001$ , and for black-box discriminator networks  $\lambda_{\text{discrim}} = 0.001$ . (Note that the critic/discriminator networks are

updated more greedily). Per usual in soft actor-critic algorithms, we also employ a lagged target critic network (i.e. used for bootstrapping); this is updated using polyak averaging at a rate of  $\tau = 0.005$ .

**Performance Metrics** We use the original source code for computing the TSTR metric (i.e. Predictive Score), publicly available at: <https://github.com/jsyoon0823/TimeGAN>; this is straightforwardly modified to compute TSTR scores for horizons of lengths three (+3 Steps Ahead) and five (+5 Steps Ahead). Likewise, we use the original source code for computing the cross-correlation score ( $x$ -Corr. Score), this is also publicly available at: <https://github.com/tianlinxu312/cot-gan>.

## D Clarifying the Analogy

This section clarifies the analogy of “generation as imitation”. While the point of our investigation is to explicitly invite an analogy between synthetic time-series generation and imitation learning, they are different problems with different considerations. In particular, TimeGCI is not suitable for imitation learning per se, and we make no claim that it is (Appendix D.1). Conversely, while existing adversarial imitation learning methods may be naively applied to time-series generation, their optimization objectives are different, and we empirically verify they do not perform well (Appendix D.2).

### D.1 Can TimeGCI be used for Imitation Learning?

At its core, time-series generation is the problem of modeling a distribution of trajectories  $p(\tau)$  faithfully, and is what TimeGCI does. In MDP parlance, in time-series generation the “environment dynamics” are by construction *known* and *deterministic*:  $\omega(\cdot|z_t, u_t)$  is the Dirac delta centered at  $z_{t+1} = (u_1, \dots, u_t)$ . We have  $p(\tau) = \prod_t \pi(u_t|z_t)$ . However, in imitation learning the dynamics are generally *unknown* and *stochastic*. We have  $p(\tau) = \prod_t \pi(u_t|z_t)\omega(z_t|z_{t-1}, u_{t-1})$ . This difference is crucial. Consider walking through the logic of Section 3, but applying it to imitation learning instead. Begin with the gradient update for learning the Gibbs parameters (Equation 7). To avoid costly inner-loop optimization of  $\theta$  to completion, we again consider (1) importance sampling, as well as our preferred method of (2) contrastive estimation. We will see that in imitation learning neither of these work out of the box. For (1), the importance weights now become  $\exp(F_\phi(\tau))/p_\theta(\tau)$ , where  $p_\theta(\tau) = \prod_t \pi_\theta(u_t|z_t)\omega(z_t|z_{t-1}, u_{t-1})$ . But the problem is that we do not have access to  $\omega$ . We can naively *estimate* it beforehand (clearly inadvisable). Or, we can perform a specific *approximation* in the energy model:  $p_\phi(\tau) = \exp(F_\phi(\tau) - \log Z_\phi) \approx \frac{1}{Z_{\phi,\omega}} \exp(F_\phi(\tau)) \prod_t \omega(z_t|z_{t-1}, u_{t-1})$  such that  $\omega$  cancels out from the expression. Instead of modeling the distribution of trajectories  $p(\tau)$  exactly, this now assumes that transition randomness has a limited effect on behavior and that the partition function is constant for all random outcome samples (see e.g. [67]). The situation is similar for (2), where it is easy to see that in imitation learning scenarios Equation 10 is no longer accessible without performing the above approximation to cancel out  $\omega$ . Three points bear emphasis. First, if we are willing to make this simplification, we are no longer speaking of TimeGCI anymore. Instead, it is easy to show that this effectively becomes a sort of trajectory-centric “maximum *causal* entropy” inverse optimal control (see e.g. [39]), which is no longer learning  $p(\tau)$  with an exact objective. Second, in general we have little reason to believe that such a trajectory-centric approach would work well in imitation learning: The variance of sample-based estimates is now exacerbated by the unknown and stochastic environment dynamics in imitation learning, such that trajectory-based sampling is generally known to perform quite poorly (see e.g. [54]). In fact, for this reason modern adversarial imitation learning methods almost always take a transition-centric approach. Third, empirically a similar structured-classifier approach (with the above approximation) has indeed been found to be unworkable in practice due to high variance (see [54] Section 4, implementing [53]); moreover, an importance-sampling based approach (with the above approximation) has been made to work, but has required hand-crafted, domain-specific regularization to work, and the learned energy functions only explain the demonstrations locally [68]. For these reasons, we do not evaluate TimeGCI on imitation learning scenarios. It is simply not applicable without modifying it to become a different method altogether. Modern imitation learning methods are designed specifically to handle the unknown and stochastic nature of environment transitions. In particular, transition-centric approaches (i.e. scoring  $(z, u)$  pairs) have been found to be more effective empirically. What we are doing with TimeGCI, on the other hand, is to focus on an exact objective for learning time-series trajectories  $p(\tau)$ , which requires a trajectory-centric approach. This is because this allows us to equate TimeGCI with a special kind of noise-contrastive estimation, for which we can enjoy some theoretical guarantees. (Note that these properties are lost when performing the above approximation in an imitation learning setting).

## D.2 Can AIL be used for Time-series Generation?

Table 4: Performance Comparison of GAIL, AIRL, and TimeGCI. Bold numbers indicate best-performing results.

Benchmark	Metric	Sines	Energy	Gas	Metro	MIMIC-III
GAIL-SAC	Predictive Score	0.447 ± 0.002	0.261 ± 0.001	0.022 ± 0.002	0.257 ± 0.001	0.017 ± 0.001
	+3 Steps Ahead	0.472 ± 0.003	0.262 ± 0.001	0.048 ± 0.001	0.250 ± 0.001	0.012 ± 0.001
	+5 Steps Ahead	0.542 ± 0.004	0.261 ± 0.001	0.074 ± 0.002	0.245 ± 0.003	0.011 ± 0.001
	$x$ -Corr. Score	6.798 ± 0.014	142.457 ± 0.471	111.124 ± 0.232	1.044 ± 0.031	540.89 ± 0.159
AIRL-SAC	Predictive Score	0.223 ± 0.006	0.283 ± 0.002	0.0478 ± 0.0015	0.243 ± 0.001	0.018 ± 0.005
	+3 Steps Ahead	0.381 ± 0.002	0.295 ± 0.003	0.0883 ± 0.0029	0.250 ± 0.001	0.019 ± 0.003
	+5 Steps Ahead	0.349 ± 0.005	0.321 ± 0.001	0.1176 ± 0.0040	0.251 ± 0.002	0.019 ± 0.001
	$x$ -Corr. Score	10.397 ± 0.005	202.61 ± 0.119	144.79 ± 0.3199	1.286 ± 0.098	2635.57 ± 0.050
TimeGCI	Predictive Score	<b>0.097 ± 0.001</b>	<b>0.251 ± 0.001</b>	<b>0.018 ± 0.000</b>	<b>0.239 ± 0.001</b>	<b>0.002 ± 0.000</b>
	+3 Steps Ahead	<b>0.104 ± 0.001</b>	<b>0.251 ± 0.001</b>	<b>0.042 ± 0.001</b>	<b>0.239 ± 0.001</b>	<b>0.001 ± 0.000</b>
	+5 Steps Ahead	<b>0.109 ± 0.001</b>	<b>0.251 ± 0.001</b>	<b>0.067 ± 0.001</b>	<b>0.239 ± 0.001</b>	<b>0.001 ± 0.000</b>
	$x$ -Corr. Score	<b>1.195 ± 0.011</b>	<b>105.2 ± 0.433</b>	<b>47.91 ± 0.811</b>	<b>0.738 ± 0.019</b>	<b>194.3 ± 0.180</b>

Conversely to the preceding, we can also ask: Can adversarial imitation learning methods work for time-series generation (hence can we compare them against TimeGCI)? The answer is “yes”, they can be *applied* to time-series generation, but “no”, there is no reason to expect they would perform better. Modern adversarial imitation learning methods come in two broad flavors: (1) GAIL-based [107–110], which do not recover a reward function, and (2) AIRL-based [54, 69, 111], which do recover a reward function. Both are transition-centric (i.e. scoring  $(z, u)$  pairs). For (1): Here, optimization is of the familiar GAN-like objective:  $\min_{\theta} \max_{\phi} \mathbb{E}_{z, u \sim \mu_s} \log d_{\phi}(z, u) + \mathbb{E}_{z, u \sim \mu_{\theta}} \log(1 - d_{\phi}(z, u))$ , which minimizes the JS-divergence between the *state-action* occupancy measures (i.e. distribution of transitions, in time-series generation terms) induced by the learned and source policies. Note that this is the same as the TimeGAN objective [12], modulo entropic/supervised regularization. As noted before, matching the transition marginals is indirectly performing the “global” sort of moment-matching. However, as pertains time-series generation, the key difference between GAIL and TimeGAN is that the former performs policy optimization for the generator, whereas the latter simply performs backpropagation. Otherwise, note that both methods are adversarial (viz. saddle-point optimization) and do not learn an explicit energy (viz. black-box discriminator). In light of this, we may consider GAIL for experimental comparison (to be consistent, we also use SAC as the policy optimization method); see Table 4 above. For (2): Here, the discriminator is first constructed to be *transition*-based:  $d_{\theta, \phi}(z, u) = (\exp(g_{\phi}(z, u)) / (\exp(g_{\phi}(z, u)) + \pi_{\theta}(u|z)))$ . Note that this actually breaks the relationship with modeling the distribution of *trajectories* directly  $p_{\phi}(\tau) = \exp(F_{\phi}(\tau) - \log Z_{\phi})$ . Importantly, if we apply AIRL to time-series generation, we lose the convergence property that comes from the relationship with noise-contrastive estimation (cf. Proposition 4), and the gradient of the discriminator is no longer related to the original trajectory-wise energy-based gradient (cf. Proposition 5). In fact, all we are left with is that the global optimum is a correct optimum—but this by itself is not helpful (e.g. the global optimum of naive MLE is also a correct optimum, but there is no guarantee that it can be found effectively). In fact, it has been formally shown that the original theoretical justifications for AIRL are incorrect (see e.g. [111] Sec. 2.4.2, or [112] Sec. B.2). (In stochastic domains in imitation learning, it is true that the empirical benefit of variance-reduction from sampling in this transition-based manner still dominates. In time-series generation, however, there is no stochasticity in the “environment”, and doing it this way may unnecessarily bias our objective of learning  $p(\tau)$ ). In light of this, we may consider AIRL for experimental comparison (to be consistent, we also use SAC for policy optimization); see Table 4.

## D.3 Ablation Studies on Sequence Lengths

First, we compare the performance of T-Forcing to TimeGCI, using sequence lengths  $T \in \{2, 8, 24\}$ . We compute the TSTR predictive score, +3 steps ahead, +5 steps ahead, and  $x$ -Corr. score as usual for each of these settings. Note that for  $T = 2$  we cannot ask the TSTR evaluation model to predict more than one step ahead, since there is no data for more than one step ahead; we indicate this with “N/A”. The Gas dataset is used. For ease of interpretation, we consider T-Forcing as the “baseline”, and also express the TimeGCI numbers as a fraction of the T-Forcing numbers. The results are consistent with what we would expect: The performance advantage enjoyed by TimeGCI over T-Forcing diminishes as the sequence lengths of the input dataset decreases, and increases as the sequence lengths of the input dataset increases. This is true regardless of what metric we are talking about (i.e. 1/3/5-step TSTR score, or feature correlation score). Moreover, at  $T = 2$  there is almost no difference between

the TSTR scores of T-Forcing and TimeGCI. This shows that T-Forcing appears to preserve quite well the one-step ahead relationships of the original data (viz.  $T = 2$ ). However, for longer sequences TimeGCI performs relatively better, which is consistent with the motivation behind using an objective that matches the distribution of trajectories, instead of simply matching transition conditionals:

Benchmark	Metric	$T = 2$	$T = 8$	$T = 24$
T-Forcing	Predictive Score	$0.038 \pm 0.001$	$0.036 \pm 0.007$	$0.035 \pm 0.003$
	+3 Steps Ahead	N/A	$0.076 \pm 0.007$	$0.080 \pm 0.001$
	+5 Steps Ahead	N/A	$0.114 \pm 0.003$	$0.111 \pm 0.001$
	x-Corr. Score	$160.3 \pm 0.346$	$154.2 \pm 0.219$	$150.8 \pm 0.067$
TimeGCI	Predictive Score	$0.037 \pm 0.001$	$0.022 \pm 0.000$	$0.018 \pm 0.000$
	+3 Steps Ahead	N/A	$0.048 \pm 0.001$	$0.042 \pm 0.001$
	+5 Steps Ahead	N/A	$0.071 \pm 0.001$	$0.067 \pm 0.001$
	x-Corr. Score	$140.4 \pm 0.173$	$50.72 \pm 0.816$	$47.91 \pm 0.811$
Ratio of Mean	Predictive Score	97.37%	61.11%	51.43%
	+3 Steps Ahead	N/A	63.16%	52.50%
	+5 Steps Ahead	N/A	62.28%	60.36%
	x-Corr. Score	87.59%	32.89%	31.77%

Second, we use a new synthetic simulation where what is being generated is a set of multi-dimensional sinusoidal waves which—if not perturbed by noise—correspond to different frequencies, phases, etc. This allows us to manually inject noise where we want, and also allows us to simulate “ground truth”. First, we train both models (T-Forcing and TimeGCI) using the same training data generated by the simulator, and save these learned models. Then during validation, we obtain a sequence from the simulator, then add some additional noise. Specifically, we add independent Gaussian noise  $\mathcal{N}(\mu, \sigma)$ , using  $\mu = 0$  and  $\sigma = 0.1$ , to each of the feature dimensions of the wave at time  $K$ . Then, our task is to predict  $t$  steps ahead with the learned models. To be clear, we ask the learned models to predict ahead using the information up to and including this (perturbed)  $K$ -th step; if the prediction is for  $t > 1$ , we perform open-loop sampling as usual. We can measure the model’s prediction error (i.e. evaluating how well the model forecasts the future based on the provided histories as input, along with the perturbation). Note that this is entirely different from the TSTR predictive score hitherto used (i.e. evaluating how well the model’s freely generated output dataset preserves the characteristics of the input dataset); this is only done as an additional sensitivity analysis: By analogy to imitation learning, we can interpret this experiment as a special case of “action-matching” (i.e. based on ground-truth state/history as input, we can measure the “prediction error” between the action chosen by the imitator policy and the actual action taken by the expert)—but specifically where the state/history has been perturbed with additional noise. We ask each model to perform  $t$ -steps ahead prediction, where  $t \in \{1, 2, 3, 4, 5\}$  as a sensitivity. When we add noise, we use values  $\{\sigma, 2\sigma, 3\sigma, 4\sigma, 5\sigma\}$  as a sensitivity. Given any sequence, the step  $K$  at which noise is artificially added (and at which point the models are asked to perform  $t$ -step ahead predictions) is uniformly randomly sampled from  $K \in \{1, 2, \dots, T - 1\}$ .

Noise at $K$	Benchmark	1-Ahead MSE	2-Ahead MSE	3-Ahead MSE	4-Ahead MSE	5-Ahead MSE
$\sigma$	T-Forcing	$0.024 \pm 0.000$	$0.029 \pm 0.001$	$0.036 \pm 0.001$	$0.041 \pm 0.001$	$0.047 \pm 0.001$
	TimeGCI	$0.024 \pm 0.000$	$0.025 \pm 0.001$	$0.026 \pm 0.001$	$0.028 \pm 0.001$	$0.031 \pm 0.001$
$2\sigma$	T-Forcing	$0.055 \pm 0.000$	$0.059 \pm 0.001$	$0.065 \pm 0.001$	$0.071 \pm 0.002$	$0.076 \pm 0.001$
	TimeGCI	$0.055 \pm 0.001$	$0.055 \pm 0.001$	$0.057 \pm 0.001$	$0.059 \pm 0.000$	$0.061 \pm 0.001$
$3\sigma$	T-Forcing	$0.106 \pm 0.001$	$0.109 \pm 0.002$	$0.116 \pm 0.001$	$0.121 \pm 0.001$	$0.127 \pm 0.002$
	TimeGCI	$0.106 \pm 0.001$	$0.106 \pm 0.001$	$0.107 \pm 0.001$	$0.109 \pm 0.001$	$0.111 \pm 0.001$
$4\sigma$	T-Forcing	$0.176 \pm 0.001$	$0.181 \pm 0.002$	$0.185 \pm 0.001$	$0.191 \pm 0.002$	$0.197 \pm 0.002$
	TimeGCI	$0.176 \pm 0.002$	$0.176 \pm 0.002$	$0.177 \pm 0.001$	$0.179 \pm 0.002$	$0.181 \pm 0.001$
$5\sigma$	T-Forcing	$0.266 \pm 0.003$	$0.271 \pm 0.002$	$0.277 \pm 0.002$	$0.280 \pm 0.002$	$0.287 \pm 0.002$
	TimeGCI	$0.266 \pm 0.002$	$0.267 \pm 0.003$	$0.267 \pm 0.002$	$0.270 \pm 0.003$	$0.272 \pm 0.003$

The results give some orthogonal intuition as to why the proposed method is better; specifically, is it because of better first-step prediction, or is it because of better robustness when having small errors in the previous steps? Observe that for *single*-step prediction, there is virtually no difference between the performance of T-Forcing and TimeGCI when evaluating the prediction MSE. The more noise added, the worse both models perform; but there is little difference between them, no matter the noise. On the other hand, for *multi*-step prediction, when predicting multiple steps ahead using open-loop sampling, T-Forcing performs worse than TimeGCI. In fact, the gap between their performances increases as  $t$  increases. So it appears that it is not the case that TimeGCI simply has better first-step prediction per se; also, it appears that TimeGCI does have better robustness when having errors in the previous steps. Because, while the 1-step performance of both T-Forcing and TimeGCI are impacted almost equally, TimeGCI appears to have an advantage in terms of “propagating” less of that error into later time steps.

## References

- [1] Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 2018.
- [2] Anna L Buczak, Steven Babin, and Linda Moniz. Data-driven approach for creating synthetic electronic medical records. *BMC medical informatics and decision making*, 2010.
- [3] Saloni Dash, Andrew Yale, Isabelle Guyon, and Kristin P Bennett. Medical time-series data generation using generative adversarial networks. *International Conference on Artificial Intelligence in Medicine (AIME)*, 2020.
- [4] James Jordon, Daniel Jarrett, Evgeny Saveliev, Jinsung Yoon, Paul Elbers, Patrick Thorat, Ari Ercole, Cheng Zhang, Danielle Belgrave, and Mihaela van der Schaar. Hide-and-seek privacy challenge: Synthetic data generation vs. patient re-identification. *NeurIPS 2020 Competition and Demonstration Track*, 2021.
- [5] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1989.
- [6] Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *International Conference on Learning Representations (ICLR)*, 2016.
- [7] Yoshua Bengio and Paolo Frasconi. An input output hmm architecture. *Advances in neural information processing systems (NeurIPS)*, 1995.
- [8] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. *International Conference on Machine Learning (ICML)*, 2009.
- [9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research (JMLR)*, 2016.
- [10] Alex Lamb, Anirudh Goyal, Ying Zhang, Saizheng Zhang, Aaron Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
- [11] Ferenc Huszár. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *International Conference on Learning Representations (ICLR)*, 2016.
- [12] Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series generative adversarial networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [13] Anirudh Goyal, Alessandro Sordoni, Marc-Alexandre Côté, Nan Rosemary Ke, and Yoshua Bengio. Z-forcing: Training stochastic recurrent networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
- [14] Ahmed M Alaa, Alex J Chan, and Mihaela van der Schaar. Generative time-series modeling with fourier flows. *International Conference on Learning Representations (ICLR)*, 2020.
- [15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [16] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint*, 2014.
- [17] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

- [18] Olof Mogren. C-rnn-gan: Continuous recurrent neural networks with adversarial training. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
- [19] Zinan Lin, Alankar Jain, Chen Wang, Giulia Fanti, and Vyas Sekar. Generating high-fidelity, synthetic time series datasets with doppelganger. *ACM Internet Measurement Conference (IMC)*, 2019.
- [20] Tianlin Xu, Li K Wenliang, Michael Munn, and Beatrice Acciaio. Cot-gan: Generating sequential data via causal optimal transport. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [21] Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint*, 2017.
- [22] Giorgia Ramponi, Pavlos Protopoulos, Marco Brambilla, and Ryan Janssen. T-cgan: Conditional generative adversarial network for data augmentation in noisy time series with irregular sampling. *arXiv preprint*, 2018.
- [23] Luca Simonetto. Generating spiking time series with generative adversarial networks: an application on banking transactions. 2018.
- [24] Moustafa Alzantot, Supriyo Chakraborty, and Mani Srivastava. Sensegen: A deep learning architecture for synthetic sensor data generation. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 188–193. IEEE, 2017.
- [25] Shota Haradal, Hideaki Hayashi, and Seichi Uchida. Biosignal data augmentation based on generative adversarial networks. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 368–371. IEEE, 2018.
- [26] Chi Zhang, Sanmukh R Kuppannagari, Rajgopal Kannan, and Viktor K Prasanna. Generative adversarial network for synthetic time series data generation in smart grids. In *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pages 1–6. IEEE, 2018.
- [27] Ahmed M Alaa, Boris van Breugel, Evgeny Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. *International Conference on Machine Learning (ICML)*, 2021.
- [28] Aditya Grover, Jiaming Song, Alekh Agarwal, Kenneth Tran, Ashish Kapoor, Eric Horvitz, and Stefano Ermon. Bias correction of learned generative models using likelihood-free importance weighting. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [29] Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. *International Conference on Learning Representations (ICLR)*, 2019.
- [30] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. Gansynth: Adversarial neural audio synthesis. *International Conference on Learning Representations (ICLR)*, 2019.
- [31] Weili Nie, Nina Narodytska, and Ankit Patel. Relgan: Relational generative adversarial networks for text generation. *International Conference on Learning Representations (ICLR)*, 2019.
- [32] Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. Language gans falling short. *International Conference on Learning Representations (ICLR)*, 2020.
- [33] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. *IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [34] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

- [35] Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. An actor-critic algorithm for sequence prediction. *International Conference on Learning Representations (ICLR)*, 2017.
- [36] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. *Advances in Neural Information Processing Systems (NeurIPS)*, 2000.
- [37] Peter D Grünwald, A Philip Dawid, et al. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *Annals of Statistics*, 2004.
- [38] Farzan Farnia and David Tse. A minimax approach to supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
- [39] Brian D Ziebart. Modeling purposeful adaptive behavior with the principle of maximum causal entropy. *Dissertation, Carnegie Mellon University*, 2010.
- [40] Michael I Jordan. An introduction to probabilistic graphical models. 2003.
- [41] Taesup Kim and Yoshua Bengio. Deep directed generative models with energy-based probability estimation. *International Conference on Learning Representations (ICLR)*, 2016.
- [42] Shuangfei Zhai, Yu Cheng, Rogerio Feris, and Zhongfei Zhang. Generative adversarial networks as variational training of energy based models. *arXiv preprint*, 2016.
- [43] Zihang Dai, Amjad Almahairi, Philip Bachman, Eduard Hovy, and Aaron Courville. Calibrating energy-based generative adversarial networks. *International Conference on Learning Representations (ICLR)*, 2017.
- [44] Rithesh Kumar, Sherjil Ozair, Anirudh Goyal, Aaron Courville, and Yoshua Bengio. Maximum entropy generators for energy-based models. *arXiv preprint*, 2019.
- [45] Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2016.
- [46] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. *International Conference on Machine Learning (ICML)*, 2017.
- [47] Wenjie Shi, Shiji Song, and Cheng Wu. Soft policy gradient method for maximum entropy deep reinforcement learning. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.
- [48] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- [49] Josiah P Hanna and Peter Stone. Towards a data efficient off-policy policy gradient. *AAAI Symposium on Data Efficient Reinforcement Learning (AAAI)*, 2018.
- [50] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Unsupervised as supervised learning. *The Elements of Statistical Learning*, 2009.
- [51] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.
- [52] Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research (JMLR)*, 2012.
- [53] Chelsea Finn, Paul Christiano, Pieter Abbeel, and Sergey Levine. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *NeurIPS Workshop on Adversarial Training*, 2016.
- [54] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *International Conference on Learning Representations (ICLR)*, 2018.

- [55] Ian J Goodfellow. On distinguishability criteria for estimating generative models. *International Conference on Learning Representations (ICLR)*, 2015.
- [56] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. Deep learning. *MIT Press Cambridge*, 2016.
- [57] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *International Conference on Machine Learning (ICML)*, 2018.
- [58] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. *International conference on artificial intelligence and statistics (AISTATS)*, 2011.
- [59] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 2018.
- [60] Alexandre Attia and Sharone Dayan. Global overview of imitation learning. *arXiv preprint*, 2018.
- [61] Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. *International conference on artificial intelligence and statistics (AISTATS)*, 2010.
- [62] Umar Syed and Robert E Schapire. A reduction from apprenticeship learning to classification. *Advances in neural information processing systems (NeurIPS)*, 2010.
- [63] Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. *International conference on Machine learning (ICML)*, 2000.
- [64] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. *International conference on Machine learning (ICML)*, 2004.
- [65] Ioana Bica, Daniel Jarrett, Alihan Hüyük, and Mihaela van der Schaar. Learning what-if explanations for sequential decision-making. *International Conference on Learning Representations (ICLR)*, 2021.
- [66] Daniel Jarrett, Alihan Hüyük, and Mihaela Van Der Schaar. Inverse decision modeling: Learning interpretable representations of behavior. *International Conference on Machine Learning (ICML)*, 2021.
- [67] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. *AAAI Conference on Artificial Intelligence (AAAI)*, 2008.
- [68] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. *International conference on machine learning (ICML)*, 2016.
- [69] Ahmed H Qureshi, Byron Boots, and Michael C Yip. Adversarial imitation via variational inverse reinforcement learning. *International Conference on Learning Representations (ICLR)*, 2019.
- [70] Ilya Kostrikov, Kumar Krishna Agrawal, Debidatta Dwibedi, Sergey Levine, and Jonathan Tompson. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation. *International Conference on Learning Representations (ICLR)*, 2019.
- [71] Lionel Blondé and Alexandros Kalousis. Sample-efficient imitation learning via gans. *International conference on artificial intelligence and statistics (AISTATS)*, 2019.
- [72] Huan Xu and Shie Mannor. Distributionally robust markov decision processes. *Mathematics of Operations Research*, 2012.
- [73] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting Structured Data*, 2006.

- [74] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A theory of generative convnet. *International Conference on Machine Learning (ICML)*, 2016.
- [75] Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [76] Philip Bachman and Doina Precup. Data generation as sequential decision making. *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.
- [77] Arun Venkatraman, Martial Hebert, and J Bagnell. Improving multi-step prediction of learned time series models. *AAAI Conference on Artificial Intelligence (AAAI)*, 2015.
- [78] Yaser Keneshloo, Tian Shi, Naren Ramakrishnan, and Chandan K Reddy. Deep reinforcement learning for sequence-to-sequence models. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2019.
- [79] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.
- [80] Andriy Mnih and Yee Whye Teh. A fast and simple algorithm for training neural probabilistic language models. *International Conference on Machine Learning (ICML)*, 2012.
- [81] Andriy Mnih and Koray Kavukcuoglu. Learning word embeddings efficiently with noise-contrastive estimation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.
- [82] Zhuang Ma and Michael Collins. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.
- [83] Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. Globally normalized transition-based neural networks. *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016.
- [84] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. *AAAI Conference on Artificial Intelligence (AAAI)*, 2017.
- [85] Sidi Lu, Lantao Yu, Siyuan Feng, Yaoming Zhu, and Weinan Zhang. Cot: Cooperative training for generative modeling of discrete data. *International Conference on Machine Learning (ICML)*, 2019.
- [86] Haiyan Yin, Dingcheng Li, Xu Li, and Ping Li. Meta-cotgan: A meta cooperative training paradigm for improving adversarial text generation. *AAAI Conference on Artificial Intelligence (AAAI)*, 2020.
- [87] William Fedus, Ian Goodfellow, and Andrew M Dai. Maskgan: better text generation via filling in the\_. *International Conference on Learning Representations (ICLR)*, 2018.
- [88] Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. Long text generation via adversarial training with leaked information. *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.
- [89] Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. Adversarial feature matching for text generation. *International Conference on Machine Learning (ICML)*, 2017.
- [90] Luis M Candanedo, Véronique Feldheim, and Dominique Deramaix. Data driven prediction models of energy use of appliances in a low-energy house. *Energy and buildings*, 2017.
- [91] Javier Burgués, Juan Manuel Jiménez-Soto, and Santiago Marco. Estimation of the limit of detection in semiconductor gas sensors through linearized calibration models. *Analytica Chimica Acta*, 2018.
- [92] John Hogue. Hourly interstate 94 westbound traffic volume for mn dot atr station 301. *Minnesota Department of Transportation*, 2018.

- [93] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Nature Scientific Data*, 2016.
- [94] Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series generative adversarial networks. <https://github.com/jsyoon0823/TimeGAN>, 2019.
- [95] Tianlin Xu, Li K Wenliang, Michael Munn, and Beatrice Acciaio. Cot-gan: Generating sequential data via causal optimal transport. <https://github.com/tianlinxu312/cot-gan>, 2020.
- [96] Olof Mogren. C-rnn-gan: Continuous recurrent neural networks with adversarial training. <https://github.com/olofmogren/c-rnn-gan>, 2016.
- [97] Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. <https://github.com/ratschlab/RGAN>, 2017.
- [98] Alex Lamb, Anirudh Goyal, Ying Zhang, Saizheng Zhang, Aaron Courville, and Yoshua Bengio. Professor forcing. [https://github.com/anirudh9119/LM\\_GANS](https://github.com/anirudh9119/LM_GANS), 2016.
- [99] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [100] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *International Conference on Very Large Data Bases (VLDB)*, 2018.
- [101] Mohammad Navid Fekri, Ananda Mohon Ghosh, and Katarina Grolinger. Generating energy data for machine learning with recurrent generative adversarial networks. *Energies*, 2020.
- [102] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. *International Conference on Learning Representations (ICLR)*, 2019.
- [103] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. *International Conference on Machine Learning (ICML)*, 2002.
- [104] Alekh Agarwal, Nan Jiang, and Sham M Kakade. Reinforcement learning: Theory and algorithms. 2019.
- [105] Gokul Swamy, Sanjiban Choudhury, Steven Wu, and Andrew Bagnell. Of moments and matching tradeoffs and treatments in imitation learning. *International Conference on Machine Learning (ICML)*, 2021.
- [106] Tian Xu, Ziniu Li, and Yang Yu. Error bounds of imitating policies and environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021.
- [107] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems (NeurIPS)*, 2016.
- [108] Nir Baram, Oron Anschel, and Shie Mannor. Model-based adversarial imitation learning. *International Conference on Machine Learning (ICML)*, 2017.
- [109] Wonseok Jeon, Seokin Seo, and Kee-Eung Kim. A bayesian approach to generative adversarial imitation learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
- [110] Xin Zhang, Yanhua Li, Ziming Zhang, and Zhi-Li Zhang.  $f$ -gail: Learning  $f$ -divergence for generative adversarial imitation learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [111] Oleg Arenz and Gerhard Neumann. Non-adversarial imitation learning and its connections to adversarial methods. *arXiv preprint*, 2020.
- [112] Tianwei Ni, Harshit Sikchi, Yufei Wang, Tejus Gupta, Lisa Lee, and Benjamin Eysenbach. F-irl: Inverse reinforcement learning via state marginal matching. *Conference on Robot Learning (CoRL)*, 2020.