A. Proof of Theorem 1

Assumption 1. *The function $q_{\boldsymbol{\theta}}(\boldsymbol{x})f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))$ and its partial derivative $\nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\boldsymbol{x})f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))$ are continuous w.r.t. $\boldsymbol{\theta}$ and $\boldsymbol{x}$.*

Theorem 1. *For a $\boldsymbol{\theta}$-parametrized energy-based model $q_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))}{Z_{\boldsymbol{\theta}}}$, under Assumption 1, the gradient of the variational representation of f-divergence (Equation (3)) w.r.t. $\boldsymbol{\theta}$ can be written as:*

$$-\nabla_{\boldsymbol{\theta}}\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))] = \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))] - \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})] \cdot \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))]$$

When we can obtain i.i.d. samples from $q_{\boldsymbol{\theta}}$, we can get an unbiased estimation of the gradient.

Proof. Under Assumption 1, with Leibniz integral rule, we have:

$$\begin{aligned}
-\nabla_{\boldsymbol{\theta}}\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))] &= -\int \nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\boldsymbol{x}) f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \\
&= -\int q_{\boldsymbol{\theta}}(\boldsymbol{x}) \frac{\nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\boldsymbol{x})}{q_{\boldsymbol{\theta}}(\boldsymbol{x})} f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \\
&= -\int q_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\boldsymbol{x}) f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \\
&= -\int q_{\boldsymbol{\theta}}(\boldsymbol{x}) \left[-\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) - \nabla_{\boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}}\right] f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \\
&= \int q_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} + \nabla_{\boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}} \cdot \int q_{\boldsymbol{\theta}}(\boldsymbol{x}) f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \\
&= \int q_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} + \frac{\int \exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))(-\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}}{Z_{\boldsymbol{\theta}}} \cdot \int q_{\boldsymbol{\theta}}(\boldsymbol{x}) f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \\
&= \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))] - \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})] \cdot \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))]
\end{aligned}$$

We can simply use $i.i.d.$ samples from $q_{\boldsymbol{\theta}}$ to get an unbiased estimation of the first term, and we introduce the following lemma for estimating the second term $\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})] \cdot \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[f^*(T_{\boldsymbol{\omega}}(\boldsymbol{x}))]$.

Lemma 3. *Let P denote a probability distribution over sample space $\mathcal{X}$ and $g, h : \mathcal{X} \to \mathbb{R}$ are functions. The following empirical estimator is unbiased with respect to $\mathbb{E}_P[g(X)] \cdot \mathbb{E}_P[h(X)]$:*

$$\frac{1}{nm}\sum_{i=1}^{n} g(x_i) \cdot \sum_{i=1}^{m} h(y_j) \tag{21}$$

where $\boldsymbol{x}_{1:n} \sim P$, $\boldsymbol{y}_{1:m} \sim P$ are two independent sets of i.i.d. samples from P.

Proof. Let $X_{1:n}$ and $Y_{1:m}$ denote the random variables corresponding to the sampling process for $\boldsymbol{x}_{1:n}$, $\boldsymbol{y}_{1:m}$ respectively. Let P_n and P_m denote the probability distribution for $X_{1:n}$ and $Y_{1:m}$. Then we have:

$$\mathbb{E}_{P_n}\mathbb{E}_{P_m}\left[\frac{1}{nm}\sum_{i=1}^{n} g(X_i) \cdot \sum_{i=1}^{m} h(Y_j)\right] \tag{22}$$

$$= \mathbb{E}_{P_n}\left[\frac{1}{n}\sum_{i=1}^{n} g(X_i) \cdot \mathbb{E}_{P_m}\left[\frac{1}{m}\sum_{i=1}^{m} h(Y_j)\right]\right] \tag{23}$$

$$= \mathbb{E}_{P_n}\left[\frac{1}{n}\sum_{i=1}^{n} g(X_i) \cdot \mathbb{E}_P[h(X)]\right] \tag{24}$$

$$= \mathbb{E}_P[g(X)] \cdot \mathbb{E}_P[h(X)] \tag{25}$$

where the first equality comes from the law of total expectation and the fact that $X_{1:n}$ and $Y_{1:m}$ are independent. Therefore, the estimator is unbiased. □

Lemma 3 shows that we can use $i.i.d.$ samples from $q_{\boldsymbol{\theta}}$ to get an unbiased estimation of the second term. □

B. Proof of Theorem 2

Theorem 2. *Let P and Q be two probability measures over the Borel σ-algebra on domain $\mathcal{X}$ with densities p, q and $P \ll Q$. Additionally, let q be an energy-based distribution, $q(\boldsymbol{x}) = \exp(-E(\boldsymbol{x}))/Z_q$. For any class of functions $\mathcal{H}$ mapping from $\mathcal{X}$ to $\mathbb{R}$ such that the expectations in the following equation are finite and $\forall \boldsymbol{x} \in \mathcal{X}$, $\mathcal{H}$ contains an element $H(\boldsymbol{x}) = \log\left(p(\boldsymbol{x})Z_q\right)$, then we have*

$$D_f(P\|Q) = \sup_{H\in\mathcal{H}} \mathbb{E}_{p(\boldsymbol{x})}[f'(\exp(H(\boldsymbol{x}) + E(\boldsymbol{x})))] - \mathbb{E}_{q(\boldsymbol{x})}[f^*(f'(\exp(H(\boldsymbol{x}) + E(\boldsymbol{x}))))]$$

where the supreme is attained at $H^*(\boldsymbol{x}) = \log\left(p(\boldsymbol{x})Z_q\right)$.

Proof. First, we define a function class $\hat{\mathcal{T}} = \{f'(r) | r : \mathcal{X} \to \mathbb{R}_+\}$. Since $\hat{\mathcal{T}}$ is a subset of $\mathcal{T}$ (in Lemma 1) and the optimal $T^\star = f'(p/q)$ is an element of $\hat{\mathcal{T}}$, from Lemma 1, we have:

$$D_f(P\|Q) = \sup_{T\in\hat{\mathcal{T}}} \mathbb{E}_{p(\boldsymbol{x})}[T(\boldsymbol{x})] - \mathbb{E}_{q(\boldsymbol{x})}[f^*(T(\boldsymbol{x}))]. \tag{26}$$

Without loss of generality, we can reparametrize all functions $T \in \hat{\mathcal{T}}$ with H such that $T(\boldsymbol{x}) = f'(\exp(H(\boldsymbol{x}) + E(\boldsymbol{x})))$ for all $\boldsymbol{x} \in \mathcal{X}$, since the transformations $f'(\cdot)$, $\exp(\cdot)$ and addition by $E(\boldsymbol{x})$ are all bijections (note that f is strictly convex and differentiable). The optimal $H^\star(\boldsymbol{x})$ then satisfies:

$$\forall \boldsymbol{x} \in \mathcal{X},\ f'(\exp(H^\star(\boldsymbol{x}) + E(\boldsymbol{x}))) = T^\star(\boldsymbol{x}) = f'(p(\boldsymbol{x})/q(\boldsymbol{x}))$$

From the bijectivity of $f'(\cdot)$, we have:

$$\begin{aligned} H^\star(\boldsymbol{x}) &= \log p(\boldsymbol{x}) - \log q(\boldsymbol{x}) - E(\boldsymbol{x}) \\ &= \log p(\boldsymbol{x}) + E(\boldsymbol{x}) + \log Z_q - E(\boldsymbol{x}) = \log(p(\boldsymbol{x})Z_q), \end{aligned}$$

□

C. Proof of Theorem 3

Assumption 2. *The function $p(\boldsymbol{x})f'(\exp(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})))$ and its partial derivative $\nabla_{\boldsymbol{\theta}} f'(\exp(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})))$ are continuous w.r.t. $\boldsymbol{\theta}$ and $\boldsymbol{x}$.*

Assumption 3. *The function $q_{\boldsymbol{\theta}}(\boldsymbol{x})f^*(f'(\exp(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x}))))$ and its partial derivative $\nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\boldsymbol{x})f^*(f'(\exp(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x}))))$ are continuous w.r.t. $\boldsymbol{\theta}$ and $\boldsymbol{x}$.*

Theorem 3. *For a $\boldsymbol{\theta}$-parametrized energy-based model $q_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))}{Z_{\boldsymbol{\theta}}}$ and a fixed $\boldsymbol{\omega}$, under Assumptions 2 and 3, the gradient of $\mathcal{L}_{f\text{-EBM}}(\boldsymbol{\theta}, \boldsymbol{\omega})$ (the objective in Equation (14)) with respect to $\boldsymbol{\theta}$ can be written as:*

$$\begin{aligned} \nabla_{\boldsymbol{\theta}}\mathcal{L}_{f\text{-EBM}}(\boldsymbol{\theta}, \boldsymbol{\omega}) = \mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} f'(\exp(f_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(x)))] + \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})] - \\ \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})] - \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})] \cdot \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})] \end{aligned}$$

where $F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x}) = f^*(f'(\exp(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x}))))$. When we can obtain i.i.d. samples from p and $q_{\boldsymbol{\theta}}$, we can get an unbiased estimation of the gradient.

Proof. Under Assumption 2, with Leibniz integral rule, for the first term in Equation (14), we have:

$$\nabla_{\boldsymbol{\theta}}\mathbb{E}_{p(\boldsymbol{x})}[f'(\exp(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})))] = \int p(\boldsymbol{x})\nabla_{\boldsymbol{\theta}} f'(\exp(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})))\mathrm{d}\boldsymbol{x} = \mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} f'(\exp(f_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(x)))]$$

For notational simplicity, let us use $F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})$ to denote $f^*(f'(\exp(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x}))))$. Under Assumption 3, with Leibniz integral rule, for the second term, we have:

$$
\begin{aligned}
&-\nabla_{\boldsymbol{\theta}}\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[f^*(f'(\exp(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x}))))] \\
=&-\int \nabla_{\boldsymbol{\theta}}(q_{\boldsymbol{\theta}}(\boldsymbol{x})F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} \\
=&-\int F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})\nabla_{\boldsymbol{\theta}}q_{\boldsymbol{\theta}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int q_{\boldsymbol{\theta}}(\boldsymbol{x})\nabla_{\boldsymbol{\theta}}F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
=&-\int q_{\boldsymbol{\theta}}(\boldsymbol{x})F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})\frac{\nabla_{\boldsymbol{\theta}}q_{\boldsymbol{\theta}}(\boldsymbol{x})}{q_{\boldsymbol{\theta}}(\boldsymbol{x})}\mathrm{d}\boldsymbol{x} - \int q_{\boldsymbol{\theta}}(\boldsymbol{x})\nabla_{\boldsymbol{\theta}}F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
=&-\int q_{\boldsymbol{\theta}}(\boldsymbol{x})F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})\nabla_{\boldsymbol{\theta}}\log q_{\boldsymbol{\theta}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int q_{\boldsymbol{\theta}}(\boldsymbol{x})\nabla_{\boldsymbol{\theta}}F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
=&-\int q_{\boldsymbol{\theta}}(\boldsymbol{x})F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})\left[-\nabla_{\boldsymbol{\theta}}E_{\boldsymbol{\theta}}(\boldsymbol{x}) - \nabla_{\boldsymbol{\theta}}\log Z_{\boldsymbol{\theta}}\right]\mathrm{d}\boldsymbol{x} - \int q_{\boldsymbol{\theta}}(\boldsymbol{x})\nabla_{\boldsymbol{\theta}}F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
=&\int q_{\boldsymbol{\theta}}(\boldsymbol{x})F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})\nabla_{\boldsymbol{\theta}}E_{\boldsymbol{\theta}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int q_{\boldsymbol{\theta}}(\boldsymbol{x})\nabla_{\boldsymbol{\theta}}F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \nabla_{\boldsymbol{\theta}}\log Z_{\boldsymbol{\theta}}\cdot\int q_{\boldsymbol{\theta}}(\boldsymbol{x})F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
=&\int q_{\boldsymbol{\theta}}(\boldsymbol{x})F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})\nabla_{\boldsymbol{\theta}}E_{\boldsymbol{\theta}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int q_{\boldsymbol{\theta}}(\boldsymbol{x})\nabla_{\boldsymbol{\theta}}F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \frac{\int \exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))(-\nabla_{\boldsymbol{\theta}}E_{\boldsymbol{\theta}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}}{Z_{\boldsymbol{\theta}}}\cdot\int q_{\boldsymbol{\theta}}(\boldsymbol{x})F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
=&\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})\nabla_{\boldsymbol{\theta}}E_{\boldsymbol{\theta}}(\boldsymbol{x})] - \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}}F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})] - \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}}E_{\boldsymbol{\theta}}(\boldsymbol{x})]\cdot\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[F_{\boldsymbol{\theta},\boldsymbol{\omega}}(\boldsymbol{x})]
\end{aligned}
$$

Similar to the proof in Appendix A, we can use *i.i.d.* samples from p and $q_{\boldsymbol{\theta}}$ to get an unbiased estimation of the gradient. □

D. Proof for the Local Convergence of f-EBM

D.1. Non-linear Dynamical Systems

In this section, we present a brief introduction of non-linear dynamical system theory (Hassan, 1996). For a comprehensive description, please refer to (Hassan, 1996; Nagarajan & Kolter, 2017). We further generalize some of the theories which will be useful for establishing the local convergence property of f-EBM later.

Consider a system consisting of variables $\boldsymbol{\phi} \in \Phi \subseteq \mathbb{R}^n$ whose time derivative is defined by the vector field $v(\boldsymbol{\phi})$:

$$\dot{\boldsymbol{\phi}} = v(\boldsymbol{\phi}) \tag{27}$$

where $v : \Phi \to \mathbb{R}^n$ is a locally Lipschitz mapping from a domain Φ into $\mathbb{R}^n$.

Suppose $\boldsymbol{\phi}^*$ is an equilibrium point of the system in Equation (27), *i.e.*, $v(\boldsymbol{\phi}^*) = 0$. Let $\boldsymbol{\phi}_t$ denote the state of the system at time t. To begin with, we introduce the following definition to characterize the stability of $\boldsymbol{\phi}^*$:

Definition 3 (Definition 4.1 in (Hassan, 1996))**.** *The equilibrium point $\boldsymbol{\phi}^*$ for the system defined in Equation (27) is*

- *stable if for each $\epsilon > 0$, there is $\delta = \delta(\epsilon) > 0$ such that*

$$\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^*\| < \delta \Longrightarrow \forall t \geq 0,\ \|\boldsymbol{\phi}_t - \boldsymbol{\phi}^*\| < \epsilon.$$

- *unstable if not stable.*

- *asymptotically stable if it is stable and $\delta > 0$ can be chosen such that*

$$\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^*\| < \delta \Longrightarrow \lim_{t\to\infty}\boldsymbol{\phi}_t = \boldsymbol{\phi}^*.$$

- *exponentially stable if it is asymptotically stable and $\delta, k, \lambda > 0$ can be chosen such that:*

$$\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^*\| < \delta \Longrightarrow \|\boldsymbol{\phi}_t\| \leq k\|\boldsymbol{\phi}_0\|\exp(-\lambda t).$$

The system is stable if for any value of ϵ, we can find a value of δ (possibly dependent on ϵ), such that a trajectory starting in a δ neighborhood of the equilibrium point will never leave the ϵ neighborhood of the equilibrium point. However, such a system may either converge to the equilibrium point or orbit within the ϵ ball. By contrast, asymptotic stability is a stronger notion of stability in the sense that trajectories starting in a δ neighborhood of the equilibrium will converge to the equilibrium point in the limit $t \to \infty$. Moreover, if $\boldsymbol{\phi}^*$ is asymptotically stable, we call the algorithm obtained by iteratively applying the updates in Equation (27) *locally convergent* to $\boldsymbol{\phi}^*$. If $\boldsymbol{\phi}^*$ is exponentially stable, we call the corresponding algorithm *linearly convergent* to $\boldsymbol{\phi}^*$.

Now we introduce the following theorem, which is central for studying the asymptotic stability of a system:

Theorem 4 (Theorem 4.15 in (Hassan, 1996))**.** *Let $\boldsymbol{\phi}^*$ be an equilibrium point for the non-linear system*

$$\dot{\boldsymbol{\phi}} = v(\boldsymbol{\phi}) \tag{28}$$

where $v : \Phi \to \mathbb{R}^n$ is continuously differentiable and Φ is a neighborhood of $\boldsymbol{\phi}^*$. Let $\boldsymbol{J}$ be the Jacobian of the system in Equation (28) at the equilibrium point:

$$\boldsymbol{J} = \left.\frac{\partial v(\boldsymbol{\phi})}{\partial \boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*} \tag{29}$$

Then, we have:

- *The equilibrium point $\boldsymbol{\phi}^*$ is asymptotically stable and exponentially stable if $\boldsymbol{J}$ is a Hurwitz matrix, i.e., $\mathrm{Re}(\lambda) < 0$ for all eigenvalues λ of $\boldsymbol{J}$.*
- *The equilibrium point $\boldsymbol{\phi}^*$ is unstable if $\mathrm{Re}(\lambda) > 0$ for one or more of the eigenvalues of $\boldsymbol{J}$.*

Proof. See proof for Theorem 4.7, Theorem 4.15 and Corollary 4.3 in (Hassan, 1996). □

With Theorem 4, the stability of an equilibrium point can be analyzed by examining if all the eigenvalues of the Jacobian $v'(\boldsymbol{\phi})|_{\boldsymbol{\phi}=\boldsymbol{\phi}^*}$ have strictly negative real part.

In the following, we will use $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ to denote the largest and smallest eigenvalues of a non-zero positive semi-definite matrix. Now we introduce the following theorem to upper bound the real part of the eigenvalues of a matrix with a specific form:

Theorem 5. *Suppose $\boldsymbol{J} \in \mathbb{R}^{(m+n)\times(m+n)}$ is of the following form:*

$$\boldsymbol{J} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{P} \\ -\boldsymbol{P}^\top & -\boldsymbol{Q} \end{bmatrix}$$

where $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ is a symmetric real positive definite matrix and $\boldsymbol{P}^\top \in \mathbb{R}^{n\times m}$ is a full column rank matrix. Then, for every eigenvalue λ of $\boldsymbol{J}$, $\mathrm{Re}(\lambda) < 0$. More precisely, we have:

- *When $\mathrm{Im}(\lambda) = 0$,*

$$\mathrm{Re}(\lambda) \le -\frac{\lambda_{\min}(\boldsymbol{Q})\lambda_{\min}(\boldsymbol{P}\boldsymbol{P}^\top)}{\lambda_{\min}(\boldsymbol{Q})\lambda_{\max}(\boldsymbol{Q}) + \lambda_{\min}(\boldsymbol{P}\boldsymbol{P}^\top)}$$

- *When $\mathrm{Im}(\lambda) \neq 0$,*

$$\mathrm{Re}(\lambda) \le -\frac{\lambda_{\min}(\boldsymbol{Q})}{2}$$

Proof. See Lemma G.2 in (Nagarajan & Kolter, 2017). □

This theorem is useful for proving the local convergence of GANs. To prove the stability of f-EBM in the following sections, we need the following generalized theorem:

Theorem 6. *Suppose $\boldsymbol{J} \in \mathbb{R}^{(m+n)\times(m+n)}$ is of the following form:*

$$\boldsymbol{J} = \begin{bmatrix} -\boldsymbol{S} & \boldsymbol{P} \\ -\boldsymbol{P}^\top & -\boldsymbol{Q} \end{bmatrix}$$

where $\boldsymbol{S} \in \mathbb{R}^{m\times m}$ is a symmetric real positive semi-definite matrix, $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ is a symmetric real positive definite matrix, $\boldsymbol{P}^\top \in \mathbb{R}^{n\times m}$ is a full column rank matrix. Then, for every eigenvalue λ of $\boldsymbol{J}$, $\mathrm{Re}(\lambda) < 0$. More precisely, we have:

- *When $\mathrm{Im}(\lambda) = 0$,*

$$\lambda_1 < -\frac{\lambda_{\min}(\boldsymbol{Q})\lambda_{\min}(\boldsymbol{P}\boldsymbol{P}^\top)}{\lambda_{\min}(\boldsymbol{Q})\lambda_{\max}(\boldsymbol{S},\boldsymbol{Q}) + \lambda_{\min}(\boldsymbol{P}\boldsymbol{P}^\top)} < 0$$

  where $\lambda_{\max}(\boldsymbol{S},\boldsymbol{Q}) = \max(\lambda_{\max}(\boldsymbol{S}), \lambda_{\max}(\boldsymbol{Q}))$.

- *When $\mathrm{Im}(\lambda) \neq 0$,*

$$\lambda_1 \leq -\frac{\lambda_{\min}(\boldsymbol{S}) + \lambda_{\min}(\boldsymbol{Q})}{2} < 0$$

Proof. We prove this theorem in a similar way to the proof for Theorem 5. Consider the following generic eigenvector equation:

$$\begin{bmatrix} -\boldsymbol{S} & \boldsymbol{P} \\ -\boldsymbol{P}^\top & -\boldsymbol{Q} \end{bmatrix} \begin{bmatrix} \boldsymbol{a}_1 + i\boldsymbol{a}_2 \\ \boldsymbol{b}_1 + i\boldsymbol{b}_2 \end{bmatrix} = (\lambda_1 + i\lambda_2) \begin{bmatrix} \boldsymbol{a}_1 + i\boldsymbol{a}_2 \\ \boldsymbol{b}_1 + i\boldsymbol{b}_2 \end{bmatrix}$$

where $\boldsymbol{a}_i, \boldsymbol{b}_i, \lambda_i$ are all real valued and the vector is normalized, *i.e.*, $\|\boldsymbol{a}_1\|^2 + \|\boldsymbol{a}_2\|^2 + \|\boldsymbol{b}_1\|^2 + \|\boldsymbol{b}_2\|^2 = 1$. The above equation can be rewritten as:

$$\begin{bmatrix} -\boldsymbol{S}\boldsymbol{a}_1 + \boldsymbol{P}\boldsymbol{b}_1 + i(-\boldsymbol{S}\boldsymbol{a}_2 + \boldsymbol{P}\boldsymbol{b}_2) \\ -\boldsymbol{P}^\top\boldsymbol{a}_1 - \boldsymbol{Q}\boldsymbol{b}_1 + i(-\boldsymbol{P}^\top\boldsymbol{a}_2 - \boldsymbol{Q}\boldsymbol{b}_2) \end{bmatrix} = \begin{bmatrix} \lambda_1\boldsymbol{a}_1 - \lambda_2\boldsymbol{a}_2 + i(\lambda_1\boldsymbol{a}_2 + \lambda_2\boldsymbol{a}_1) \\ \lambda_1\boldsymbol{b}_1 - \lambda_2\boldsymbol{b}_2 + i(\lambda_1\boldsymbol{b}_2 + \lambda_2\boldsymbol{b}_1) \end{bmatrix}$$

By equating the real and complex parts, we get:

$$-\boldsymbol{S}\boldsymbol{a}_1 + \boldsymbol{P}\boldsymbol{b}_1 = \lambda_1\boldsymbol{a}_1 - \lambda_2\boldsymbol{a}_2 \tag{30}$$
$$-\boldsymbol{P}^\top\boldsymbol{a}_1 - \boldsymbol{Q}\boldsymbol{b}_1 = \lambda_1\boldsymbol{b}_1 - \lambda_2\boldsymbol{b}_2 \tag{31}$$
$$-\boldsymbol{S}\boldsymbol{a}_2 + \boldsymbol{P}\boldsymbol{b}_2 = \lambda_1\boldsymbol{a}_2 + \lambda_2\boldsymbol{a}_1 \tag{32}$$
$$-\boldsymbol{P}^\top\boldsymbol{a}_2 - \boldsymbol{Q}\boldsymbol{b}_2 = \lambda_1\boldsymbol{b}_2 + \lambda_2\boldsymbol{b}_1 \tag{33}$$

By multiplying the Equations (30), (31), (32), (33) by $\boldsymbol{a}_1^\top, \boldsymbol{b}_1^\top, \boldsymbol{a}_2^\top, \boldsymbol{b}_2^\top$ and adding them together, we get:

$$\begin{aligned} &-\boldsymbol{a}_1^\top\boldsymbol{S}\boldsymbol{a}_1 + \boldsymbol{a}_1^\top\boldsymbol{P}\boldsymbol{b}_1 - \boldsymbol{b}_1^\top\boldsymbol{P}^\top\boldsymbol{a}_1 - \boldsymbol{b}_1^\top\boldsymbol{Q}\boldsymbol{b}_1 - \boldsymbol{a}_2^\top\boldsymbol{S}\boldsymbol{a}_2 + \boldsymbol{a}_2^\top\boldsymbol{P}\boldsymbol{b}_2 - \boldsymbol{b}_2^\top\boldsymbol{P}^\top\boldsymbol{a}_2 - \boldsymbol{b}_2^\top\boldsymbol{Q}\boldsymbol{b}_2 \\ =&\lambda_1\boldsymbol{a}_1^\top\boldsymbol{a}_1 - \lambda_2\boldsymbol{a}_1^\top\boldsymbol{a}_2 + \lambda_1\boldsymbol{b}_1^\top\boldsymbol{b}_1 - \lambda_2\boldsymbol{b}_1^\top\boldsymbol{b}_2 + \lambda_1\boldsymbol{a}_2^\top\boldsymbol{a}_2 + \lambda_2\boldsymbol{a}_2^\top\boldsymbol{a}_1 + \lambda_1\boldsymbol{b}_2^\top\boldsymbol{b}_2 + \lambda_2\boldsymbol{b}_2^\top\boldsymbol{b}_1 \end{aligned}$$

which simplifies to

$$-\boldsymbol{a}_1^\top\boldsymbol{S}\boldsymbol{a}_1 - \boldsymbol{a}_2^\top\boldsymbol{S}\boldsymbol{a}_2 - \boldsymbol{b}_1^\top\boldsymbol{Q}\boldsymbol{b}_1 - \boldsymbol{b}_2^\top\boldsymbol{Q}\boldsymbol{b}_2 = \lambda_1(\boldsymbol{a}_1^\top\boldsymbol{a}_1 + \boldsymbol{b}_1^\top\boldsymbol{b}_1 + \boldsymbol{a}_2^\top\boldsymbol{a}_2 + \boldsymbol{b}_2^\top\boldsymbol{b}_2) = \lambda_1$$

Because $S \succeq \boldsymbol{0}, Q \succ \boldsymbol{0}$, $-\boldsymbol{a}_1^\top\boldsymbol{S}\boldsymbol{a}_1 - \boldsymbol{a}_2^\top\boldsymbol{S}\boldsymbol{a}_2 - \boldsymbol{b}_1^\top\boldsymbol{Q}\boldsymbol{b}_1 - \boldsymbol{b}_2^\top\boldsymbol{Q}\boldsymbol{b}_2 \leq 0$, where the equality holds only if $\boldsymbol{b}_1 = \boldsymbol{0}, \boldsymbol{b}_2 = \boldsymbol{0}$ (as well as $-\boldsymbol{a}_1^\top\boldsymbol{S}\boldsymbol{a}_1 - \boldsymbol{a}_2^\top\boldsymbol{S}\boldsymbol{a}_2 = 0$). Next we show that $\boldsymbol{b}_1 = \boldsymbol{0}, \boldsymbol{b}_2 = \boldsymbol{0}$ is contradictory with the condition that $\boldsymbol{P}^\top$ is a full column rank matrix. First, when $-\boldsymbol{a}_1^\top\boldsymbol{S}\boldsymbol{a}_1 - \boldsymbol{a}_2^\top\boldsymbol{S}\boldsymbol{a}_2 - \boldsymbol{b}_1^\top\boldsymbol{Q}\boldsymbol{b}_1 - \boldsymbol{b}_2^\top\boldsymbol{Q}\boldsymbol{b}_2 = 0$, we have $\lambda_1 = 0$. Applying these to Equations (31) and (33), we get $\boldsymbol{P}^\top\boldsymbol{a}_1 = 0$ and $\boldsymbol{P}^\top\boldsymbol{a}_2 = 0$. Since one of $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$ is non-zero (otherwise the eigenvector is zero), this implies $\boldsymbol{P}^\top$ is not a full column rank matrix, which is contradictory with the condition that $\boldsymbol{P}^\top$ has full column rank. Consequently, $-\boldsymbol{a}_1^\top\boldsymbol{S}\boldsymbol{a}_1 - \boldsymbol{a}_2^\top\boldsymbol{S}\boldsymbol{a}_2 - \boldsymbol{b}_1^\top\boldsymbol{Q}\boldsymbol{b}_1 - \boldsymbol{b}_2^\top\boldsymbol{Q}\boldsymbol{b}_2 = \lambda_1 < 0$.

Now we proceed to get a tighter upper bound on the real part of the eigenvalue λ_1. By multiplying Equations (30) and (32) by $-\boldsymbol{a}_2^\top$ and $\boldsymbol{a}_1^\top$ and adding them together, we get:

$$\boldsymbol{a}_2^\top\boldsymbol{S}\boldsymbol{a}_1 - \boldsymbol{a}_2^\top\boldsymbol{P}\boldsymbol{b}_1 - \boldsymbol{a}_1^\top\boldsymbol{S}\boldsymbol{a}_2 + \boldsymbol{a}_1^\top\boldsymbol{P}\boldsymbol{b}_2 = -\lambda_1\boldsymbol{a}_2^\top\boldsymbol{a}_1 + \lambda_2\boldsymbol{a}_2^\top\boldsymbol{a}_2 + \lambda_1\boldsymbol{a}_1^\top\boldsymbol{a}_2 + \lambda_2\boldsymbol{a}_1^\top\boldsymbol{a}_1$$

which simplifies to:

$$-\boldsymbol{a}_2^\top \boldsymbol{P}\boldsymbol{b}_1 + \boldsymbol{a}_1^\top \boldsymbol{P}\boldsymbol{b}_2 = \lambda_2 \boldsymbol{a}_2^\top \boldsymbol{a}_2 + \lambda_2 \boldsymbol{a}_1^\top \boldsymbol{a}_1 \tag{34}$$

Similarly, by multiplying Equations (31) and (33) by $-\boldsymbol{b}_2^\top$ and $\boldsymbol{b}_1^\top$ and adding them together, we get:

$$\boldsymbol{b}_2^\top \boldsymbol{P}^\top \boldsymbol{a}_1 + \boldsymbol{b}_2^\top \boldsymbol{Q}\boldsymbol{b}_1 - \boldsymbol{b}_1^\top \boldsymbol{P}^\top \boldsymbol{a}_2 - \boldsymbol{b}_1^\top \boldsymbol{Q}\boldsymbol{b}_2 = -\lambda_1 \boldsymbol{b}_2^\top \boldsymbol{b}_1 + \lambda_2 \boldsymbol{b}_2^\top \boldsymbol{b}_2 + \lambda_1 \boldsymbol{b}_1^\top \boldsymbol{b}_2 + \lambda_2 \boldsymbol{b}_1^\top \boldsymbol{b}_1 \tag{35}$$

which simplifies to:

$$\boldsymbol{b}_2^\top \boldsymbol{P}^\top \boldsymbol{a}_1 - \boldsymbol{b}_1^\top \boldsymbol{P}^\top \boldsymbol{a}_2 = \lambda_2 \boldsymbol{b}_2^\top \boldsymbol{b}_2 + \lambda_2 \boldsymbol{b}_1^\top \boldsymbol{b}_1 \tag{36}$$

With Equation (34) and Equation (36), we have:

$$\lambda_2(\|\boldsymbol{a}_1\|^2 + \|\boldsymbol{a}_2\|^2) = \lambda_2(\|\boldsymbol{b}_1\|^2 + \|\boldsymbol{b}_2\|^2)$$

which implies that either $\|\boldsymbol{a}_1\|^2 + \|\boldsymbol{a}_2\|^2 = \|\boldsymbol{b}_1\|^2 + \|\boldsymbol{b}_2\|^2 = 1/2$ or $\lambda_2 = 0$.

In the first case ($\lambda_2 \neq 0$ and $\|\boldsymbol{a}_1\|^2 + \|\boldsymbol{a}_2\|^2 = \|\boldsymbol{b}_1\|^2 + \|\boldsymbol{b}_2\|^2 = 1/2$), from $-\boldsymbol{a}_1^\top \boldsymbol{S}\boldsymbol{a}_1 - \boldsymbol{a}_2^\top \boldsymbol{S}\boldsymbol{a}_2 - \boldsymbol{b}_1^\top \boldsymbol{Q}\boldsymbol{b}_1 - \boldsymbol{b}_2^\top \boldsymbol{Q}\boldsymbol{b}_2 = \lambda_1$, we get an upper bound:

$$\lambda_1 \leq -\frac{\lambda_{\min}(\boldsymbol{S}) + \lambda_{\min}(\boldsymbol{Q})}{2}$$

This upper bound is strictly negative since $\lambda_{\min}(\boldsymbol{S}) \geq 0$ and $\lambda_{\min}(\boldsymbol{Q}) > 0$.

Now we introduce the following lemma which is useful for deriving the upper bound of λ_1 in the second case:

Lemma 4. *Let $\boldsymbol{S} \succeq \boldsymbol{0}$ and $\boldsymbol{Q} \succeq \boldsymbol{0}$ be two real symmetric matrices. If $\boldsymbol{a}^\top \boldsymbol{S}\boldsymbol{a} + \boldsymbol{b}^\top \boldsymbol{Q}\boldsymbol{b} = c$, then $\boldsymbol{a}^\top \boldsymbol{S}^\top \boldsymbol{S}\boldsymbol{a} + \boldsymbol{b}^\top \boldsymbol{Q}^\top \boldsymbol{Q}\boldsymbol{b} \in [c \cdot \min(\lambda_{\min}(\boldsymbol{S}), \lambda_{\min}(\boldsymbol{Q})), c \cdot \max(\lambda_{\max}(\boldsymbol{S}), \lambda_{\max}(\boldsymbol{Q}))]$.*

Proof. Let $\boldsymbol{S} = \boldsymbol{U}_{\boldsymbol{S}} \boldsymbol{\Lambda}_{\boldsymbol{S}} \boldsymbol{U}_{\boldsymbol{S}}^\top$ and $\boldsymbol{Q} = \boldsymbol{U}_{\boldsymbol{Q}} \boldsymbol{\Lambda}_{\boldsymbol{Q}} \boldsymbol{U}_{\boldsymbol{Q}}^\top$ be the eigenvalue decompositions of $\boldsymbol{S}$ and $\boldsymbol{Q}$. Then, we have

$$c = \boldsymbol{a}^\top \boldsymbol{S}\boldsymbol{a} + \boldsymbol{b}^\top \boldsymbol{Q}\boldsymbol{b} = \boldsymbol{a}^\top \boldsymbol{U}_{\boldsymbol{S}} \boldsymbol{\Lambda}_{\boldsymbol{S}} \boldsymbol{U}_{\boldsymbol{S}}^\top \boldsymbol{a} + \boldsymbol{b}^\top \boldsymbol{U}_{\boldsymbol{Q}} \boldsymbol{\Lambda}_{\boldsymbol{Q}} \boldsymbol{U}_{\boldsymbol{Q}}^\top \boldsymbol{b} = \boldsymbol{x}^\top \boldsymbol{\Lambda}_{\boldsymbol{S}} \boldsymbol{x} + \boldsymbol{y}^\top \boldsymbol{\Lambda}_{\boldsymbol{Q}} \boldsymbol{y}$$

where $\boldsymbol{x} = \boldsymbol{U}_{\boldsymbol{S}}^\top \boldsymbol{a}$ and $\boldsymbol{y} = \boldsymbol{U}_{\boldsymbol{Q}}^\top \boldsymbol{b}$. Therefore, we have:

$$c = \sum_i x_i^2 \lambda_{\boldsymbol{S}}^i + \sum_j y_j^2 \lambda_{\boldsymbol{Q}}^j$$

Similarly, we have:

$$\begin{aligned}\boldsymbol{a}^\top \boldsymbol{S}^\top \boldsymbol{S}\boldsymbol{a} + \boldsymbol{b}^\top \boldsymbol{Q}^\top \boldsymbol{Q}\boldsymbol{b} =& \boldsymbol{a}^\top \boldsymbol{U}_{\boldsymbol{S}} \boldsymbol{\Lambda}_{\boldsymbol{S}} \boldsymbol{U}_{\boldsymbol{S}}^\top \boldsymbol{U}_{\boldsymbol{S}} \boldsymbol{\Lambda}_{\boldsymbol{S}} \boldsymbol{U}_{\boldsymbol{S}}^\top \boldsymbol{a} + \boldsymbol{b}^\top \boldsymbol{U}_{\boldsymbol{Q}} \boldsymbol{\Lambda}_{\boldsymbol{Q}} \boldsymbol{U}_{\boldsymbol{Q}}^\top \boldsymbol{U}_{\boldsymbol{Q}} \boldsymbol{\Lambda}_{\boldsymbol{Q}} \boldsymbol{U}_{\boldsymbol{Q}}^\top \boldsymbol{b} \\ =& \sum_i x_i^2 (\lambda_{\boldsymbol{S}}^i)^2 + \sum_j y_j^2 (\lambda_{\boldsymbol{Q}}^j)^2\end{aligned}$$

which differs from c by a multiplicative factor within $[\min(\lambda_{\min}(\boldsymbol{S}), \lambda_{\min}(\boldsymbol{Q})), \max(\lambda_{\max}(\boldsymbol{S}), \lambda_{\max}(\boldsymbol{Q}))]$. □

In the second case, the imaginary part of the eigenvalue is zero ($\lambda_2 = 0$), which implies the imaginary part of the eigenvector must also be zero, $\boldsymbol{a}_2 = \boldsymbol{b}_2 = \boldsymbol{0}$. Applying this to Equations (30), (31), (32), (33), we get:

$$\begin{aligned}-\boldsymbol{S}\boldsymbol{a}_1 + \boldsymbol{P}\boldsymbol{b}_1 =& \lambda_1 \boldsymbol{a}_1 \\ -\boldsymbol{P}^\top \boldsymbol{a}_1 - \boldsymbol{Q}\boldsymbol{b}_1 =& \lambda_1 \boldsymbol{b}_1\end{aligned}$$

Rearranging the above equations, we get:

$$\begin{aligned}\boldsymbol{P}\boldsymbol{b}_1 =& (\lambda_1 \boldsymbol{I} + \boldsymbol{S})\boldsymbol{a}_1 \\ -\boldsymbol{P}^\top \boldsymbol{a}_1 =& (\lambda_1 \boldsymbol{I} + \boldsymbol{Q})\boldsymbol{b}_1\end{aligned}$$

Squaring both sides of the equations, we get:

$$\boldsymbol{b}_1^\top \boldsymbol{P}^\top \boldsymbol{P}\boldsymbol{b}_1 = \boldsymbol{a}_1^\top(\lambda_1^2\boldsymbol{I} + 2\lambda_1\boldsymbol{S} + \boldsymbol{S}^\top\boldsymbol{S})\boldsymbol{a}_1 = \lambda_1^2\|\boldsymbol{a}_1\|^2 + 2\lambda_1\boldsymbol{a}_1^\top\boldsymbol{S}\boldsymbol{a}_1 + \boldsymbol{a}_1^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{a}_1$$
$$\boldsymbol{a}_1^\top \boldsymbol{P}\boldsymbol{P}^\top \boldsymbol{a}_1 = \boldsymbol{b}_1^\top(\lambda_1^2\boldsymbol{I} + 2\lambda_1\boldsymbol{Q} + \boldsymbol{Q}^\top\boldsymbol{Q})\boldsymbol{b}_1 = \lambda_1^2\|\boldsymbol{b}_1\|^2 + 2\lambda_1\boldsymbol{b}_1^\top\boldsymbol{Q}\boldsymbol{b}_1 + \boldsymbol{b}_1^\top\boldsymbol{Q}^\top\boldsymbol{Q}\boldsymbol{b}_1$$

Summing these two equations together, using the fact that $-\boldsymbol{a}_1^\top\boldsymbol{S}\boldsymbol{a}_1 - \boldsymbol{b}_1^\top\boldsymbol{Q}\boldsymbol{b}_1 = \lambda_1$ and $\|\boldsymbol{a}_1\|^2 + \|\boldsymbol{b}_1\|^2 = 1$, we get:

$$\begin{aligned}\boldsymbol{b}_1^\top \boldsymbol{P}^\top \boldsymbol{P}\boldsymbol{b}_1 + \boldsymbol{a}_1^\top \boldsymbol{P}\boldsymbol{P}^\top \boldsymbol{a}_1 =& \lambda_1^2(\|\boldsymbol{a}_1\|^2 + \|\boldsymbol{b}_1\|^2) - 2\lambda_1^2 + \boldsymbol{a}_1^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{a}_1 + \boldsymbol{b}_1^\top\boldsymbol{Q}^\top\boldsymbol{Q}\boldsymbol{b}_1 \\ =& -\lambda_1^2 + \boldsymbol{a}_1^\top\boldsymbol{S}^\top\boldsymbol{S}\boldsymbol{a}_1 + \boldsymbol{b}_1^\top\boldsymbol{Q}^\top\boldsymbol{Q}\boldsymbol{b}_1\end{aligned}$$

Let $\lambda_{\max}(\boldsymbol{S}, \boldsymbol{Q})$ denote $\max(\lambda_{\max}(\boldsymbol{S}), \lambda_{\max}(\boldsymbol{Q}))$. With Lemma 4, we have:

$$\boldsymbol{b}_1^\top \boldsymbol{P}^\top \boldsymbol{P}\boldsymbol{b}_1 + \boldsymbol{a}_1^\top \boldsymbol{P}\boldsymbol{P}^\top \boldsymbol{a}_1 \leq -\lambda_1^2 - \lambda_1\lambda_{\max}(\boldsymbol{S}, \boldsymbol{Q}) \tag{37}$$

Furthermore, we have:

$$-\lambda_1 = \boldsymbol{a}_1^\top\boldsymbol{S}\boldsymbol{a}_1 + \boldsymbol{b}_1^\top\boldsymbol{Q}\boldsymbol{b}_1 \geq \lambda_{\min}(\boldsymbol{Q})\|\boldsymbol{b}_1\|^2 = \lambda_{\min}(\boldsymbol{Q})(1 - \|\boldsymbol{a}_1\|^2) \Longrightarrow \|\boldsymbol{a}_1\|^2 \geq 1 + \frac{\lambda_1}{\lambda_{\min}(\boldsymbol{Q})}$$

Note that λ_1 can either satisfy $\lambda_1 \leq -\lambda_{\min}(\boldsymbol{Q})$ or $-\lambda_{\min}(\boldsymbol{Q}) < \lambda_1 < 0$. In the first scenario, we already obtain an upper bound: $\lambda_1 \leq -\lambda_{\min}(\boldsymbol{Q})$. So we focus on deriving an upper bound for the second scenario. From Equation (37), we have:

$$\begin{aligned}&-\lambda_1^2 - \lambda_1\lambda_{\max}(\boldsymbol{S}, \boldsymbol{Q}) \geq \boldsymbol{a}_1^\top \boldsymbol{P}\boldsymbol{P}^\top \boldsymbol{a}_1 \geq \lambda_{\min}(\boldsymbol{P}\boldsymbol{P}^\top)\|\boldsymbol{a}_1\|^2 \geq \lambda_{\min}(\boldsymbol{P}\boldsymbol{P}^\top)\left(1 + \frac{\lambda_1}{\lambda_{\min}(\boldsymbol{Q})}\right) \\ \Longrightarrow &-\lambda_1\left(\lambda_1 + \lambda_{\max}(\boldsymbol{S}, \boldsymbol{Q}) + \frac{\lambda_{\min}(\boldsymbol{P}\boldsymbol{P}^\top)}{\lambda_{\min}(\boldsymbol{Q})}\right) \geq \lambda_{\min}(\boldsymbol{P}\boldsymbol{P}^\top) \\ \Longrightarrow &-\lambda_1\left(\lambda_{\max}(\boldsymbol{S}, \boldsymbol{Q}) + \frac{\lambda_{\min}(\boldsymbol{P}\boldsymbol{P}^\top)}{\lambda_{\min}(\boldsymbol{Q})}\right) > \lambda_{\min}(\boldsymbol{P}\boldsymbol{P}^\top)\end{aligned} \tag{38}$$

where the last implication uses the fact that $\lambda_1 < 0$. From Equation (38), we get an upper bound for λ_1:

$$\lambda_1 < -\lambda_{\min}(\boldsymbol{Q})\frac{\lambda_{\min}(\boldsymbol{P}\boldsymbol{P}^\top)}{\lambda_{\min}(\boldsymbol{Q})\lambda_{\max}(\boldsymbol{S}, \boldsymbol{Q}) + \lambda_{\min}(\boldsymbol{P}\boldsymbol{P}^\top)}$$

Since the fraction in above equation lies in $(0, 1)$, we will use this as the upper bound for the second case ($\lambda_2 = 0$). Also note that this upper bound is strictly negative, since $\lambda_{\max}(\boldsymbol{S}, \boldsymbol{Q}) \geq \lambda_{\min}(\boldsymbol{Q}) > 0$ and $\lambda_{\min}(\boldsymbol{P}\boldsymbol{P}^\top) > 0$.

□

D.2. Notations and Setup

First, we can reformulate the minimax game defined in Equation (14) as:

$$\min_{\boldsymbol{\theta}}\max_{\boldsymbol{\omega}} V(\boldsymbol{\theta}, \boldsymbol{\omega}) = \min_{\boldsymbol{\theta}}\max_{\boldsymbol{\omega}} \mathbb{E}_{p(\boldsymbol{x})}[A(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x}))] - \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[B(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x}))] \tag{39}$$

where the functions $A(u)$ and $B(u)$ are defined as:

$$A(u) = f'(\exp(u)),\ B(u) = f^*(f'(\exp(u))) - f^*(f'(1)) \tag{40}$$

with $B(0) = 0$. Since $f^*(f'(1))$ is a constant, and $\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[B(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x}))] = \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[f^*(f'(\exp(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x}))))] - f^*(f'(1))$, the above formulation is equivalent to the original f-EBM minimax game in Equation (14) (up to a constant that does not depend on $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$).

We have the following theorem to characterize the properties of functions $A(u)$ and $B(u)$:

Theorem 7. *For any f-divergence with closed, strictly convex and third-order differentiable generator function f, functions $A(u), B(u)$ defined as:*

$$A(u) = f'(\exp(u)),\ B(u) = f^*(f'(\exp(u))) - f^*(f'(1))$$

satisfy the following properties:

- $A'(0) = B'(0) = f''(1) > 0$
- $A''(0) - B''(0) = -f''(1) < 0$
- $A''(0) - B''(0) + 2B'(0) = f''(1) = B'(0) > 0$

where A', B' are first-order derivative of A, B; A'', B'' are second-order derivative of A, B.

Proof. First, we introduce the following lemma for convex conjugates and subgradients:

Lemma 5. *If f is closed and convex, then we have:*

$$y \in \partial f(x) \Longleftrightarrow x \in \partial f^*(y)$$

Proof. If $y \in \partial f(x)$, then $f^*(y) = \sup_u(yu - f(u)) = yx - f(x)$. Therefore, we have:

$$\begin{aligned} f^*(v) =& \sup_u vu - f(u) \\ \geq& vx - f(x) \\ =& x(v-y) - f(x) + xy \\ =& f^*(y) + x(v-y) \end{aligned}$$

Because this holds for all v, we have $x \in \partial f^*(y)$. The reverse implication $x \in \partial f^*(y) \Longrightarrow y \in \partial f(x)$ follows from $f^{**} = f$. □

Let us use $g(y)$ to denote the conjugate function, $g(y) = f^*(y)$. Next we have the following lemma for the gradient of conjugate:

Lemma 6. *If f is strictly convex and differentiable, then we have: $g'(y) = \arg\max_x\ (yx - f(x))$, $g'(f'(u)) = u$.*

Proof. Since f is strictly convex, x maximizes $yx - f(x)$ if and only if $y \in \partial f(x)$. With Lemma 5, we know that

$$y \in \partial f(x) \Longleftrightarrow x \in \partial f^*(y) = \{g'(y)\}$$

Therefore $g'(y) = \arg\max_x\ (yx - f(x))$. Then we have:

$$g'(f'(u)) = \arg\max_x\ f'(u)x - f(x) := \arg\max_x\ h(x)$$

Since $h'(x) = f'(u) - f'(x), h''(x) = -f''(x) < 0$, we have $\arg\max_x\ f'(u)x - f(x) = u$. □

Now we are ready to proof the properties of functions $A(u)$ and $B(u)$. Recall that $A(u) = f'(\exp(u))$, $B(u) = f^*(f'(\exp(u))) - f^*(f'(1))$, with Lemma 6, we have:

$$\begin{aligned} A'(u) =& f''(\exp(u))\exp(u) \\ A''(u) =& f'''(\exp(u))\exp(2u) + f''(\exp(u))\exp(u) \\ B'(u) =& g'(f'(\exp(u)))f''(\exp(u))\exp(u) \\ =& f''(\exp(u))\exp(2u) \\ B''(u) =& f'''(\exp(u))\exp(3u) + 2f''(\exp(u))\exp(2u) \end{aligned}$$

Then we have:

$$\begin{aligned} A'(0) =& f''(1) \\ A''(0) =& f'''(1) + f''(1) \\ B'(0) =& f''(1) \\ B''(0) =& f'''(1) + 2f''(1) \end{aligned}$$

Since f is strictly convex with $\forall x \in \mathrm{dom}(f), f''(x) > 0$, we have:

$$A'(0) = B'(0) = f''(1) > 0$$
$$A''(0) - B''(0) = -f''(1) < 0$$
$$A''(0) - B''(0) + 2B'(0) = f''(1) = B'(0) > 0$$

□

Some examples of Theorem 7 can be found in Table 3.

Name	$A'(0)$	$B'(0)$	$A''(0)$	$B''(0)$
Kullback-Leibler	1	1	0	1
Reverse Kullback-Leibler	1	1	-1	0
Pearson χ^2	2	2	2	4
Neyman χ^2	2	2	-4	-2
Squared Hellinger	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	$\frac{1}{4}$
Jensen-Shannon	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	$\frac{1}{4}$
α-divergence ($\alpha \notin \{0, 1\}$)	1	1	$\alpha - 1$	α

Table 3. Some examples of f-divergences and the corresponding first- and second-order gradient values of functions $A(u)$ and $B(u)$ (defined in Equation (40)) at the equilibrium point.

Let us use $(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)$ to denote the equilibrium point and we have the following realizability assumption:

Assumption 4 (Realizability)**.** $\exists \boldsymbol{\theta}^*, \boldsymbol{\omega}^*$, *such that* $\forall \boldsymbol{x} \in \mathrm{supp}(p)$, $q_{\boldsymbol{\theta}^*}(\boldsymbol{x}) = p(\boldsymbol{x})$ *and* $H_{\boldsymbol{\omega}^*}(\boldsymbol{x}) = \log(p(\boldsymbol{x})Z_{\boldsymbol{\theta}^*})$.

That is we assume the energy function and the variational functions are powerful enough such that at the equilibrium point, the model distribution $q_{\boldsymbol{\theta}^*}$ matches the true data distribution p and the variational function achieves the optimal form in Theorem 2. Note that this also implies:

$$\begin{aligned}\forall \boldsymbol{x} \in \mathrm{supp}(p), H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x}) &= \log(p(\boldsymbol{x})Z_{\boldsymbol{\theta}^*}) - \log(\exp(-E_{\boldsymbol{\theta}^*}(\boldsymbol{x}))) \\ &= \log(p(\boldsymbol{x})/q_{\boldsymbol{\theta}^*}(\boldsymbol{x})) = 0\end{aligned}$$

Furthermore, we have the following assumption on the energy function:

Assumption 5. $(\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}}^\top E_{\boldsymbol{\theta}}(\boldsymbol{x})] - 2\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})]\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}}^\top E_{\boldsymbol{\theta}}(\boldsymbol{x})])|_{\boldsymbol{\theta}^*}$ *is positive semi-definite.*

In one-dimension case, this assumption is saying: $\mathbb{E}_{p(x)}[(\nabla_\theta E_\theta(x))^2] - 2(\mathbb{E}_{p(x)}[\nabla_\theta E_\theta(x)])^2 \geq 0$. Intuitively, this implies the sum of energies over the entire sample space changes smoothly such that the average change of energies is small compared to the sum of the squared change of energies. For a better understanding of this assumption and its rationality, in the following, we will introduce two simple examples that satisfy this assumption.

Example 1. *Consider the following quadratic energy function:*

$$E_{\boldsymbol{\theta}}(x) = \frac{(x-\mu)^2}{2\sigma^2},\ \boldsymbol{\theta} = [\mu, \sigma]^\top,\ p = q_{\boldsymbol{\theta}} = \frac{\exp(-\frac{(x-\mu)^2}{2\sigma^2})}{\sqrt{2\pi}\sigma}$$

which corresponds to a Gaussian distribution.

The first-order gradient is:

$$\frac{\partial E_{\boldsymbol{\theta}}(x)}{\partial \mu} = \frac{x-\mu}{\sigma^2},\ \frac{\partial E_{\boldsymbol{\theta}}(x)}{\partial \sigma} = -\frac{(x-\mu)^2}{\sigma^3}$$

Then, we have:

$$\begin{aligned}
\int_{-\infty}^{+\infty} p(x)\frac{\partial E_{\boldsymbol{\theta}}(x)}{\partial \mu}dx &= \int_{-\infty}^{+\infty} \frac{\exp(-\frac{(x-\mu)^2}{2\sigma^2})}{\sqrt{2\pi}\sigma}\frac{x-\mu}{\sigma^2}dx = 0\\
\int_{-\infty}^{+\infty} p(x)\frac{\partial E_{\boldsymbol{\theta}}(x)}{\partial \sigma}dx &= -\int_{-\infty}^{+\infty} \frac{\exp(-\frac{(x-\mu)^2}{2\sigma^2})}{\sqrt{2\pi}\sigma}\frac{(x-\mu)^2}{\sigma^3}dx\\
&= \frac{1}{\sqrt{2\pi}\sigma^4}\int_{-\infty}^{+\infty}\exp(-\frac{x^2}{2\sigma^2})x^2dx = \frac{1}{\sigma}
\end{aligned}$$

$$\int_{-\infty}^{+\infty} p(x)\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(x)dx = \int_{-\infty}^{+\infty} p(x)\begin{bmatrix}\frac{\partial E_{\boldsymbol{\theta}}(x)}{\partial \mu}\\ \frac{\partial E_{\boldsymbol{\theta}}(x)}{\partial \sigma}\end{bmatrix}dx = \begin{bmatrix}0\\ \frac{1}{\sigma}\end{bmatrix}$$

$$\begin{aligned}
\int_{-\infty}^{+\infty} p(x)\left(\frac{\partial E_{\boldsymbol{\theta}}(x)}{\partial \mu}\right)^2 dx &= \int_{-\infty}^{+\infty} \frac{\exp(-\frac{(x-\mu)^2}{2\sigma^2})}{\sqrt{2\pi}\sigma}\frac{(x-\mu)^2}{\sigma^4}dx\\
&= \frac{1}{\sqrt{2\pi}\sigma^5}\int_{-\infty}^{+\infty}\exp(-\frac{x^2}{2\sigma^2})x^2dx = \frac{1}{\sigma^2}\\
\int_{-\infty}^{+\infty} p(x)\frac{\partial E_{\boldsymbol{\theta}}(x)}{\partial \mu}\frac{\partial E_{\boldsymbol{\theta}}(x)}{\partial \sigma}dx &= -\int_{-\infty}^{+\infty} \frac{\exp(-\frac{(x-\mu)^2}{2\sigma^2})}{\sqrt{2\pi}\sigma}\frac{(x-\mu)^3}{\sigma^5}dx = 0\\
\int_{-\infty}^{+\infty} p(x)\left(\frac{\partial E_{\boldsymbol{\theta}}(x)}{\partial \sigma}\right)^2 dx &= \int_{-\infty}^{+\infty} \frac{\exp(-\frac{(x-\mu)^2}{2\sigma^2})}{\sqrt{2\pi}\sigma}\frac{(x-\mu)^4}{\sigma^6}dx = 0\\
&= \frac{1}{\sqrt{2\pi}\sigma^7}\int_{-\infty}^{+\infty}\exp(-\frac{x^2}{2\sigma^2})x^4dx = \frac{3}{\sigma^2}
\end{aligned}$$

$$\int_{-\infty}^{+\infty} p(x)\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(x)\nabla_{\boldsymbol{\theta}}^\top E_{\boldsymbol{\theta}}(x)dx = \int_{-\infty}^{+\infty} p(x)\begin{bmatrix}\left(\frac{\partial E_{\boldsymbol{\theta}}(x)}{\partial \mu}\right)^2 & \frac{\partial E_{\boldsymbol{\theta}}(x)}{\partial \mu}\frac{\partial E_{\boldsymbol{\theta}}(x)}{\partial \sigma}\\ \frac{\partial E_{\boldsymbol{\theta}}(x)}{\partial \mu}\frac{\partial E_{\boldsymbol{\theta}}(x)}{\partial \sigma} & \left(\frac{\partial E_{\boldsymbol{\theta}}(x)}{\partial \sigma}\right)^2\end{bmatrix}dx = \begin{bmatrix}\frac{1}{\sigma^2} & 0\\ 0 & \frac{3}{\sigma^2}\end{bmatrix}$$

Therefore, Assumption 5 holds:

$$\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})\nabla_{\boldsymbol{\theta}}^\top E_{\boldsymbol{\theta}}(\boldsymbol{x})] - 2\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})]\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}}^\top E_{\boldsymbol{\theta}}(\boldsymbol{x})] = \begin{bmatrix}\frac{1}{\sigma^2} & 0\\ 0 & \frac{1}{\sigma^2}\end{bmatrix} \succ \mathbf{0}$$

Example 2. *We consider a more powerful energy function:*

$$E_{\boldsymbol{\theta}}(\boldsymbol{x}) = -\log\left(\sum_{k=1}^K \pi_k \frac{\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k))}{\sqrt{(2\pi)^n|\Sigma_k|}}\right),\ \sum_{k=1}^K \pi_k = 1,\ \boldsymbol{x}\in\mathbb{R}^n,\ \boldsymbol{\mu}_k\in\mathbb{R}^n,\ \Sigma_k \in S_{++}^n$$

where $\pi_1,\dots,\pi_K,\boldsymbol{\mu}_1,\dots,\boldsymbol{\mu}_K,\Sigma_1,\dots,\Sigma_K$ are the learnable parameters $\boldsymbol{\theta}$; S_{++}^n denotes the space of symmetric positive definite $n\times n$ matrices.

The partition function is:

$$Z_{\boldsymbol{\theta}} = \int \exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} = \int \sum_{k=1}^K \pi_k \frac{\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k))}{\sqrt{(2\pi)^n|\Sigma_k|}}\mathrm{d}\boldsymbol{x} = \sum_{k=1}^K \pi_k = 1$$

Therefore, Assumption 5 holds:

$$\begin{aligned}
&\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})\nabla_{\boldsymbol{\theta}}^\top E_{\boldsymbol{\theta}}(\boldsymbol{x})] - 2\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})]\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}}^\top E_{\boldsymbol{\theta}}(\boldsymbol{x})]\\
=&\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})\nabla_{\boldsymbol{\theta}}^\top E_{\boldsymbol{\theta}}(\boldsymbol{x})] - 2\nabla_{\boldsymbol{\theta}}\log Z_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}^\top \log Z_{\boldsymbol{\theta}}\\
=&\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})\nabla_{\boldsymbol{\theta}}^\top E_{\boldsymbol{\theta}}(\boldsymbol{x})]
\end{aligned}$$

which is a positive semi-definite moment matrix.

Note that the model distribution induced by the energy function correspond to a Gaussian mixture model (GMM). With the universal approximation theorem of GMM (i.e., any smooth density can be approximated with any specific nonzero amount of error by a GMM with enough components (Goodfellow et al., 2016)), the above EBM can be used to fit any probability distribution when K is large enough.

Finally, the last assumption we need is:

Assumption 6. $(\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \nabla_{\boldsymbol{\omega}}^\top H_{\boldsymbol{\omega}}(\boldsymbol{x})])|_{\boldsymbol{\omega}^*}$ *and* $(\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x})] \mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}}^\top E_{\boldsymbol{\theta}}(\boldsymbol{x})])|_{(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)}$ *are full column rank.*

This is similar to the assumption used in the local convergence analysis of GANs (Nagarajan & Kolter, 2017; Mescheder et al., 2018). Alternatively, we can replace it with another assumption that the rank deficiencies of the above matrices, if any, correspond to equivalent equilibria. In the following, we will use Assumption 6.

D.3. Local Convergence of Single-Step f-EBM

In the following, we will establish a theoretical proof of the local convergence property of Single-Step f-EBM. More specifically, instead of assuming the variational function $H_{\boldsymbol{\omega}}$ is optimal at every update of the energy function, we consider a more realistic setting, where both the variational function $H_{\boldsymbol{\omega}}$ and energy function $E_{\boldsymbol{\theta}}$ take simultaneous gradient steps, with time derivatives defined as:

$$\begin{pmatrix} \dot{\boldsymbol{\theta}} \\ \dot{\boldsymbol{\omega}} \end{pmatrix} = \begin{pmatrix} -\nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\omega}) \\ \nabla_{\boldsymbol{\omega}} V(\boldsymbol{\theta}, \boldsymbol{\omega}) \end{pmatrix} \tag{41}$$

That is we use gradient descent to update $\boldsymbol{\theta}$ and gradient ascent to update $\boldsymbol{\omega}$ with respect to $V(\boldsymbol{\theta}, \boldsymbol{\omega})$ at the same frequency. In practice we can also update $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ alternatively as presented in Algorithm 1. Note that our theoretical analysis also holds for the alternative gradient methods, as the ordinary differential equations of both simultaneous gradient methods and alternative gradient methods have the same Jacobians at the equilibrium point.

Throughout this paper we will use the notation $\nabla^\top(\cdot)$ to denote the row vector corresponding to the gradient that is being compute. First, we derive the Jacobian at equilibrium.

Theorem 8. *For the dynamical system defined in Equation (39) and the updates defined in Equation (41), under Assumption 4, the Jacobian at an equilibrium point $(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)$ is:*

$$\boldsymbol{J} = \begin{bmatrix} -\nabla_{\boldsymbol{\theta}}^2 V & -\nabla_{\boldsymbol{\omega}} \nabla_{\boldsymbol{\theta}} V \\ \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\omega}} V & \nabla_{\boldsymbol{\omega}}^2 V \end{bmatrix} = \begin{bmatrix} -f''(1) \boldsymbol{K}_{EE} & f''(1) \boldsymbol{K}_{EH} \\ -f''(1) \boldsymbol{K}_{EH}^\top & -f''(1) \boldsymbol{K}_{HH} \end{bmatrix}$$

where

$$\begin{aligned} \boldsymbol{K}_{EE} :=& (\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}}^\top E_{\boldsymbol{\theta}}(\boldsymbol{x})] - 2 \mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})] \mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}}^\top E_{\boldsymbol{\theta}}(\boldsymbol{x})])|_{\boldsymbol{\theta}^*} \\ \boldsymbol{K}_{EH} :=& (\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})] \mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\omega}}^\top H_{\boldsymbol{\omega}}(\boldsymbol{x})])|_{(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)} \\ \boldsymbol{K}_{HH} :=& (\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \nabla_{\boldsymbol{\omega}}^\top H_{\boldsymbol{\omega}}(\boldsymbol{x})])|_{\boldsymbol{\omega}^*} \end{aligned}$$

Proof. First, let us derive the first- and second-order derivatives of $V(\boldsymbol{\theta}, \boldsymbol{\omega})$ with respect to $\boldsymbol{\theta}$:

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\omega}) = & \int p(\boldsymbol{x}) A'(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\ & \int B(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})) \nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\ & \int q_{\boldsymbol{\theta}}(\boldsymbol{x}) B'(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \end{aligned}$$

$$\begin{aligned}\nabla^2_{\boldsymbol{\theta}} V(\boldsymbol{\theta},\boldsymbol{\omega}) = &\int p(\boldsymbol{x}) A''(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \\ &\int p(\boldsymbol{x}) A'(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})) \nabla^2_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\ &\int B'(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})) \nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\ &\int B(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})) \nabla^2_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\ &\int B'(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\ &\int q_{\boldsymbol{\theta}}(\boldsymbol{x}) B''(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\ &\int q_{\boldsymbol{\theta}}(\boldsymbol{x}) B'(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})) \nabla^2_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}\end{aligned}$$

At the equilibrium point $(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)$, we have:

$$\begin{aligned}\nabla^2_{\boldsymbol{\theta}} V(\boldsymbol{\theta},\boldsymbol{\omega})|_{(\boldsymbol{\theta}^*,\boldsymbol{\omega}^*)} = \Bigg(&\int p(\boldsymbol{x}) A''(0) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \\ &\int p(\boldsymbol{x}) A'(0) \nabla^2_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\ &\int B'(0) \nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\ &\int B(0) \nabla^2_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\ &\int B'(0) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\ &\int q_{\boldsymbol{\theta}}(\boldsymbol{x}) B''(0) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\ &\int q_{\boldsymbol{\theta}}(\boldsymbol{x}) B'(0) \nabla^2_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \Bigg) \Bigg|_{\boldsymbol{\theta}^*}\end{aligned}$$

With Theorem 7 and Assumption 4, as well as by definition of function $B(u)$ ($B(0) = 0$) we have:

$$\begin{aligned}\nabla^2_{\boldsymbol{\theta}} V(\boldsymbol{\theta},\boldsymbol{\omega})|_{(\boldsymbol{\theta}^*,\boldsymbol{\omega}^*)} = \Bigg(&\int p(\boldsymbol{x}) A''(0) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\ &\int B'(0) \nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\ &\int B'(0) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\ &\int q_{\boldsymbol{\theta}}(\boldsymbol{x}) B''(0) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \Bigg) \Bigg|_{\boldsymbol{\theta}^*}\end{aligned} \tag{42}$$

Next, for a $\boldsymbol{\theta}$-parametrized energy based model $q_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))}{Z_{\boldsymbol{\theta}}}$, we observe that

$$\nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\boldsymbol{x}) = q_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} \log q_{\boldsymbol{\theta}}(\boldsymbol{x}) = q_{\boldsymbol{\theta}}(\boldsymbol{x})(-\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) - \nabla_{\boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}}) \tag{43}$$

$$\nabla_{\boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}} = \frac{\int \exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))(-\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})) \mathrm{d}\boldsymbol{x}}{Z_{\boldsymbol{\theta}}} = -\int q_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \tag{44}$$

With Assumption 4 ($q_{\boldsymbol{\theta}^*} = p$) and above observation, Equation (42) can be written as:

$$\begin{aligned}
\nabla^2_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\omega})|_{(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)} = &\Bigg(\int p(\boldsymbol{x})(A''(0) - B''(0)) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\
&\int B'(0) \nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\
&\int B'(0) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \Bigg) \Bigg|_{\boldsymbol{\theta}^*} \\
= &\Bigg((A''(0) - B''(0)) \int p(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\
&\int B'(0) q_{\boldsymbol{\theta}}(\boldsymbol{x}) (-\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) - \nabla_{\boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}}) \nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\
&\int B'(0) q_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) (-\nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) - \nabla^\top_{\boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}}) \mathrm{d}\boldsymbol{x} \Bigg) \Bigg|_{\boldsymbol{\theta}^*} \\
= &\Bigg((A''(0) - B''(0) + 2B'(0)) \int p(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \\
&B'(0) \nabla_{\boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}} \cdot \int q_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \\
&B'(0) \int q_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \cdot \nabla^\top_{\boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}} \Bigg) \Bigg|_{\boldsymbol{\theta}^*} \\
= &\Bigg((A''(0) - B''(0) + 2B'(0)) \int p(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\
&2B'(0) \nabla_{\boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}} \nabla^\top_{\boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}} \Bigg) \Bigg|_{\boldsymbol{\theta}^*}
\end{aligned}$$

With Theorem 7, we have:

$$\begin{aligned}
\nabla^2_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\omega})|_{(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)} = &f''(1) \left(\int p(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - 2 \nabla_{\boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}} \nabla^\top_{\boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}} \right) \Bigg|_{\boldsymbol{\theta}^*} \\
= &f''(1) \left(\int p(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - 2 \int p(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \int p(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right) \Bigg|_{\boldsymbol{\theta}^*} \\
= &f''(1) (\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})] - 2 \mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})] \mathbb{E}_{p(\boldsymbol{x})}[\nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})]) |_{\boldsymbol{\theta}^*}
\end{aligned}$$

Now let us derive the first- and second-order derivatives of $V(\boldsymbol{\theta}, \boldsymbol{\omega})$ with respect to $\boldsymbol{\omega}$:

$$\nabla_{\boldsymbol{\omega}} V(\boldsymbol{\theta}, \boldsymbol{\omega}) = \int p(\boldsymbol{x}) A'(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})) \nabla_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int q_{\boldsymbol{\theta}}(\boldsymbol{x}) B'(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})) \nabla_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

$$\begin{aligned}
\nabla^2_{\boldsymbol{\omega}} V(\boldsymbol{\theta}, \boldsymbol{\omega}) = &\int p(\boldsymbol{x}) A''(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})) \nabla_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \\
&\int p(\boldsymbol{x}) A'(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})) \nabla^2_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\
&\int q_{\boldsymbol{\theta}}(\boldsymbol{x}) B''(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})) \nabla_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\
&\int q_{\boldsymbol{\theta}}(\boldsymbol{x}) B'(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})) \nabla^2_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}
\end{aligned}$$

At the equilibrium point $(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)$, under Assumption 4, we have:

$$\nabla^2_{\boldsymbol{\omega}} V(\boldsymbol{\theta}, \boldsymbol{\omega})|_{(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)} = \left((A''(0) - B''(0)) \int p(\boldsymbol{x}) \nabla_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + (A'(0) - B'(0)) \int p(\boldsymbol{x}) \nabla^2_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right) \Bigg|_{\boldsymbol{\omega}^*}$$

With Theorem 7, we have:

$$\begin{aligned} \nabla^2_{\boldsymbol{\omega}} V(\boldsymbol{\theta}, \boldsymbol{\omega})|_{(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)} &= -f''(1) \left(\int p(\boldsymbol{x}) \nabla_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right) \Bigg|_{\boldsymbol{\omega}^*} \\ &= -f''(1) (\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x})])|_{\boldsymbol{\omega}^*} \end{aligned}$$

Finally, let us derive $\nabla_{\boldsymbol{\omega}} \nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\omega})$:

$$\begin{aligned} \nabla_{\boldsymbol{\omega}} \nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\omega}) = &\int p(\boldsymbol{x}) A''(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\ &\int B'(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})) \nabla_{\boldsymbol{\theta}} q_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\ &\int q_{\boldsymbol{\theta}}(\boldsymbol{x}) B''(H_{\boldsymbol{\omega}}(\boldsymbol{x}) + E_{\boldsymbol{\theta}}(\boldsymbol{x})) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \end{aligned}$$

At the equilibrium point $(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)$, under Assumption 4, we have:

$$\begin{aligned} \nabla_{\boldsymbol{\omega}} \nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\omega})|_{(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)} &= \Bigg((A''(0) - B''(0)) \int p(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\ &\qquad B'(0) \int q_{\boldsymbol{\theta}}(\boldsymbol{x}) (-\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) - \nabla_{\boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}}) \nabla^\top_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \Bigg) \Bigg|_{(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)} \\ &= \Bigg((A''(0) - B''(0) + B'(0)) \int p(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \\ &\qquad B'(0) \nabla_{\boldsymbol{\theta}} \log Z_{\boldsymbol{\theta}} \int p(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \Bigg) \Bigg|_{(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)} \\ &= \Bigg((A''(0) - B''(0) + B'(0)) \int p(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \\ &\qquad B'(0) \int p(\boldsymbol{x}) \nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \int p(\boldsymbol{x}) \nabla^\top_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \Bigg) \Bigg|_{(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)} \end{aligned}$$

With Theorem 7, we have:

$$\nabla_{\boldsymbol{\omega}} \nabla_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, \boldsymbol{\omega})|_{(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)} = -f''(1) (\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})] \mathbb{E}_{p(\boldsymbol{x})}[\nabla^\top_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x})])|_{(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)}$$

Similarly, for $\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\omega}} V(\boldsymbol{\theta}, \boldsymbol{\omega})$, we have:

$$\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\omega}} V(\boldsymbol{\theta}, \boldsymbol{\omega})|_{(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)} = -f''(1) (\mathbb{E}_{p(\boldsymbol{x})}[\nabla_{\boldsymbol{\omega}} H_{\boldsymbol{\omega}}(\boldsymbol{x})] \mathbb{E}_{p(\boldsymbol{x})}[\nabla^\top_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})])|_{(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)}$$

□

Now we are ready to present the main theorem:

Theorem 9. *The dynamical system defined in Equation (39) and the updates defined in Equation (41) is locally exponentially stable with respect to an equilibrium point $(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)$ when the Assumptions 4, 5, 6 hold for $(\boldsymbol{\theta}^*, \boldsymbol{\omega}^*)$. Let $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the largest and smallest eigenvalues of a non-zero positive semi-definite matrix. The rate of convergence is governed only by the eigenvalues λ of the Jacobian $\boldsymbol{J}$ of the system at the equilibrium point, with a strictly negative real part upper bounded as:*

- *When* $\mathrm{Im}(\lambda) = 0$,

$$\mathrm{Re}(\lambda) < -f''(1)\frac{\lambda_{\min}(\boldsymbol{K}_{HH})\lambda_{\min}(\boldsymbol{K}_{EH}\boldsymbol{K}_{EH}^\top)}{\lambda_{\min}(\boldsymbol{K}_{HH})\lambda_{\max}(\boldsymbol{K}_{EE},\boldsymbol{K}_{HH}) + \lambda_{\min}(\boldsymbol{K}_{EH}\boldsymbol{K}_{EH}^\top)} < 0$$

  where $\lambda_{\max}(\boldsymbol{K}_{EE},\boldsymbol{K}_{HH}) = \max(\lambda_{\max}(\boldsymbol{K}_{EE}),\lambda_{\max}(\boldsymbol{K}_{HH}))$.

- *When* $\mathrm{Im}(\lambda) \neq 0$,

$$\mathrm{Re}(\lambda) \leq -\frac{f''(1)}{2}(\lambda_{\min}(\boldsymbol{K}_{EE}) + \lambda_{\min}(\boldsymbol{K}_{HH})) < 0$$

Proof. In Theorem 8, for any f-divergences with strictly convex and differentiable generator function f, we derived the Jacobian of the system $\boldsymbol{J}$ under Assumption 4. With Assumptions 5 and 6, and the strict convexity of the function f, we know that $f''(1)\boldsymbol{K}_{EE}$ is positive semi-definite, $f''(1)\boldsymbol{K}_{HH}$ is positive definite (a full rank moment matrix with a positive multiplicative factor), and $f''(1)\boldsymbol{K}_{EH}^\top$ is full column rank. Therefore, with Theorem 6, we know that the Jacobian $\boldsymbol{J}$ is a Hurwitz matrix, (*i.e.*, all the eigenvalues of $\boldsymbol{J}$ have strictly negative real parts). Furthermore, with Theorem 6, we can obtain an upper bound of the real parts of the eigenvalues. Finally, with Theorem 4, we can conclude that the system is locally exponentially stable. □

E. Implementation Details

E.1. Implementation of f-EBM Algorithm in PyTorch

```
def update_H(real_x, fake_x, model_h, model_e, optim_h, grad_exp, conjugate_grad_exp):
    // Step 6-7 in Algorithm 1. Update the parameter of the variational function.
    // - real_x and fake_x are samples from the data distribution and EBM respectively.
    // - model_h is the neural network for the variational function $H_\vomega$.
    // - model_e is the neural network for the energy function $E_\vtheta$.
    // - optim_h is the optimizer for model_h, e.g. torch.optim.SGD(model_h.parameters())
    // - grad_exp and conjugate_grad_exp are functions defined by the used f-divergence.
    real_e, fake_e = model_e(real_x), model_e(fake_x)
    real_h, fake_h = model_h(real_x), model_h(fake_x)
    loss_h = -(grad_exp(real_h+real_e) - conjugate_grad_exp(fake_h+fake_e)).mean()
    optim_h.zero_grad()
    loss_h.backward()
    optim_h.step()

def update_E(real_x, fake_x, model_h, model_e, optim_e, grad_exp, conjugate_grad_exp):
    // Step 8-9 in Algorithm 1. Update the parameter of the energy function.
    // - optim_e is the optimizer for model_e, e.g. torch.optim.SGD(model_e.parameters())
    real_e, fake_e = model_e(real_x), model_e(fake_x)
    real_h, fake_h = model_h(real_x), model_h(fake_x)
    loss_e = torch.mean(grad_exp(real_h+real_e)) + \
             torch.mean(conjugate_grad_exp(fake_h+fake_e).detach() * fake_e) - \
             torch.mean(conjugate_grad_exp(fake_h+fake_e)) - \
             torch.mean(fake_e) * torch.mean(conjugate_grad_exp(fake_h+fake_e)).detach()
    optim_e.zero_grad()
    loss_e.backward()
    optim_e.step()
```

Note that according to Lemma 3, technically we should use two independent batches of data to estimate the two expectations in the product $\mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[\nabla_{\boldsymbol{\theta}} E_{\boldsymbol{\theta}}(\boldsymbol{x})] \cdot \mathbb{E}_{q_{\boldsymbol{\theta}}(\boldsymbol{x})}[f^*(f'(\exp(E_{\boldsymbol{\theta}}(\boldsymbol{x}) + H_{\boldsymbol{\omega}}(\boldsymbol{x}))))]$. Empirically we found that using the same set of samples also works well, since it is an asymptotically consistent estimator.

To implement f-EBM in Tensorflow (Abadi et al., 2016), the main change is to use `tf.stop_gradient(X)` to replace `X.detach()`. Functions `grad_exp` and `conjugate_grad_exp` correspond to $f'(\exp(u))$ and $f^*(f'(\exp(u)))$, which

can be derived based on the definitions of f-divergences (see Table 5 and Table 6 in (Nowozin et al., 2016) for reference). Some examples of $f'(\exp(u))$ and $f^*(f'(\exp(u)))$ can be found in Table 4.

Name	$D_f(P\|Q)$	$f'(\exp(u))$	$f^*(f'(\exp(u)))$
Kullback-Leibler	$\int p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}$	$1+u$	$\exp(u)$
Reverse Kullback-Leibler	$\int q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}$	$-\exp(-u)$	$-1+u$
Pearson χ^2	$\int \frac{(q(\boldsymbol{x})-p(\boldsymbol{x}))^2}{p(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}$	$2\exp(u)-2$	$\exp(2u)-1$
Neyman χ^2	$\int \frac{(p(\boldsymbol{x})-q(\boldsymbol{x}))^2}{q(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}$	$1-\exp(-2u)$	$2-2\exp(-u)$
Squared Hellinger	$\int \left(\sqrt{p(\boldsymbol{x})}-\sqrt{q(\boldsymbol{x})}\right)^2 \mathrm{d}\boldsymbol{x}$	$1-\exp(-\frac{u}{2})$	$\exp(\frac{u}{2})-1$
Jensen-Shannon	$\frac{1}{2}\int p(\boldsymbol{x}) \log \frac{2p(\boldsymbol{x})}{p(\boldsymbol{x})+q(\boldsymbol{x})} + q(\boldsymbol{x}) \log \frac{2q(\boldsymbol{x})}{p(\boldsymbol{x})+q(\boldsymbol{x})} \mathrm{d}\boldsymbol{x}$	$\log(2)+u-\log(1+\exp(u))$	$-\log(2)+\log(1+\exp(u))$
α-divergence ($\alpha \notin \{0,1\}$)	$\frac{1}{\alpha(\alpha-1)}\int \left(p(\boldsymbol{x})\left[\left(\frac{q(\boldsymbol{x})}{p(\boldsymbol{x})}\right)^{\alpha}-1\right]-\alpha(q(\boldsymbol{x})-p(\boldsymbol{x}))\right) \mathrm{d}\boldsymbol{x}$	$\frac{1}{\alpha-1}(\exp((\alpha-1)u)-1)$	$\frac{1}{\alpha}(\exp(\alpha u)-1)$

Table 4. Some examples of f-divergences and corresponding $f'(\exp(u))$ and $f^*(f'(\exp(u)))$ functions.

E.2. Discussion on Differentiating Through Langevin Dynamics

In this section, we provide a more detailed discussion on gradient reparametrization that was initially introduced in Section 3.1.2. Specifically, we will discuss the possibility of directly extending the f-GANs framework by differentiating through the Langevin dynamics. As discussed before, we typically need hundreds of Langevin steps to produce a single sample while we cannot use a sample replay buffer to reduce the number of transition steps (because the initial distribution of the Markov chain cannot depend on the model parameters). During gradient backpropagation, we need the same number of backward steps, where each backward step further involves computing Hessian matrices that are proportional to the parameter and data dimension. This will lead to hundreds of times more memory consumption compared to only using Langevin dynamics for producing samples. More specifically, we provide the following implementation to differentiate through Langevin dynamics in PyTorch:

```
1  def train_energy(model_e, model_dis, optim_e, conjugate_fn, device='cuda'):
2      // Initialize Langevin dynamics with random uniform distribution.
3      fake_x = torch.rand(batch_size, channel_num, img_size, img_size, device=device)
4      fake_x.requires_grad = True
5      gaussian_noise = torch.randn(batch_size, 3, 32, 32, device=device)
6
7      for k in range(num_langevin_steps):
8          energy_value = model_e(fake_x)
9          fake_x_grad = torch.autograd.grad(energy_value.sum(), fake_x,
10                                            create_graph=True)[0]
11         gaussian_noise.normal_(0, gaussian_noise_std)
12         fake_x = fake_x - step_size * fake_x_grad + gaussian_noise
13         fake_x.data.clamp_(0, 1)
14
15     fake_dis = model_dis(fake_x)
16     energy_loss = - torch.mean(conjugate_fn(fake_dis))
17     optim_e.zero_grad()
18     energy_loss.backward()
19     optim_e.step()
```

Note that in Line 10, we need to set `create_graph=True` in order to compute the second-order derivatives when backpropagating through Langevin dynamics later, which will store all the computation graphs along the Langevin dynamics. By contrast, in f-EBMs we only use Langevin dynamics to produce samples and once we get the value of the gradient $\nabla_{\boldsymbol{x}} E_{\boldsymbol{\theta}}(\boldsymbol{x})$, we can discard the computational graph immediately after each transition step (`create_graph=False`). As a result, with gradient reparametrization, the memory consumption grows linearly as the number of transitions. With modern GPU such as NVIDIA GeForce RTX 2080 Ti and the model architecture in Figure 15, we will run out of memory when the number of Langevin steps is larger than 5. Since gradient backpropagation through Langevin dynamics involves computing many Hessian matrices, this approach is also computationally less efficient compared to f-EBMs. In our experiments, we set `num_langevin_steps` $= 5$ and we observed that this approach cannot produce reasonable images.

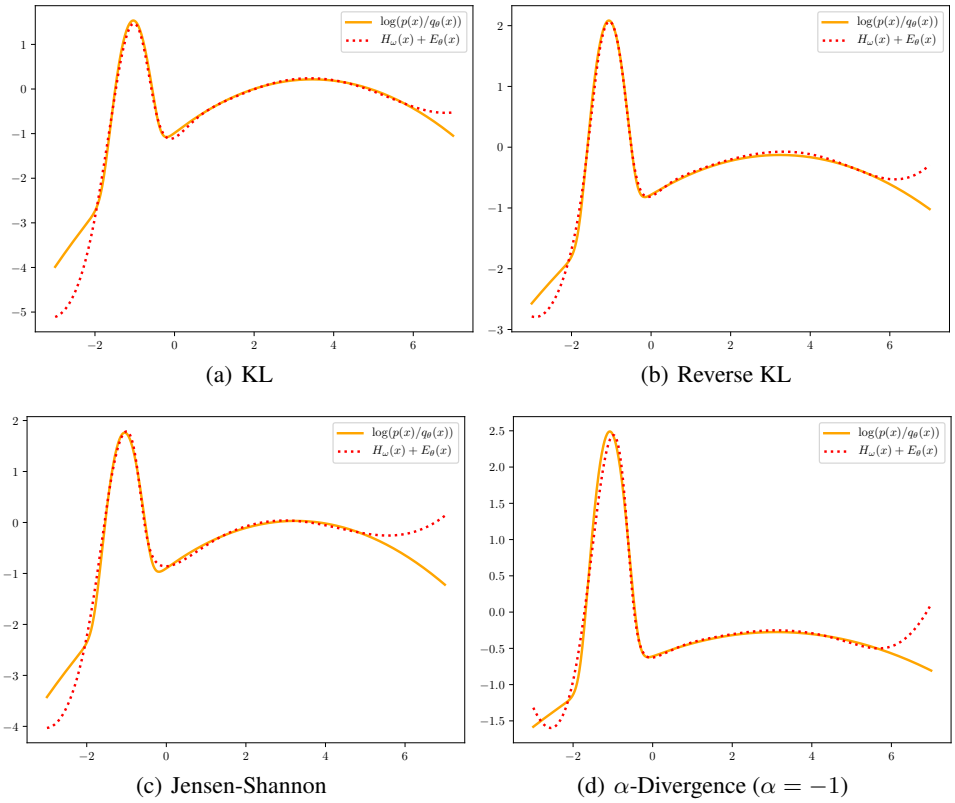
F. Additional Experimental Results for Fitting Univariate Mixture of Gaussians

F.1. Parameter Learning Results

Objective	μ^*	$\hat{\mu}$	σ^*	$\hat{\sigma}$
Contrastive Divergence	1.01065	1.01204	1.82895	1.82907
Kullback-Leibler	1.01065	1.01536	1.82895	1.83024
Reverse KL	1.58454	1.58523	1.63106	1.63453
Squared Hellinger	1.32024	1.32274	1.73089	1.74710
Jensen Shannon	1.30322	1.31669	1.76716	1.76041
Pearson χ^2	0.57581	0.56563	1.92172	1.93461
Neyman χ^2	1.83037	1.82676	1.51508	1.51598
α-divergence ($\alpha = -0.5$)	1.74642	1.74332	1.55569	1.55209
α-divergence ($\alpha = -1$)	1.82923	1.81979	1.51844	1.52513
α-divergence ($\alpha = 0.9$)	1.07056	1.07237	1.81091	1.81852

Table 5. Fitting a quadratic EBM to mixtures of Gaussians. μ^*, σ^* represent the desired optimal solution under a certain discrepancy measure, and $\hat{\mu}, \hat{\sigma}$ represent the learned parameters. The first row is for contrastive divergence (with KL divergence being the underlying objective). The other rows are for f-EBMs with various discrepancy measures as the training objectives.

F.2. Density Ratio Estimation Results

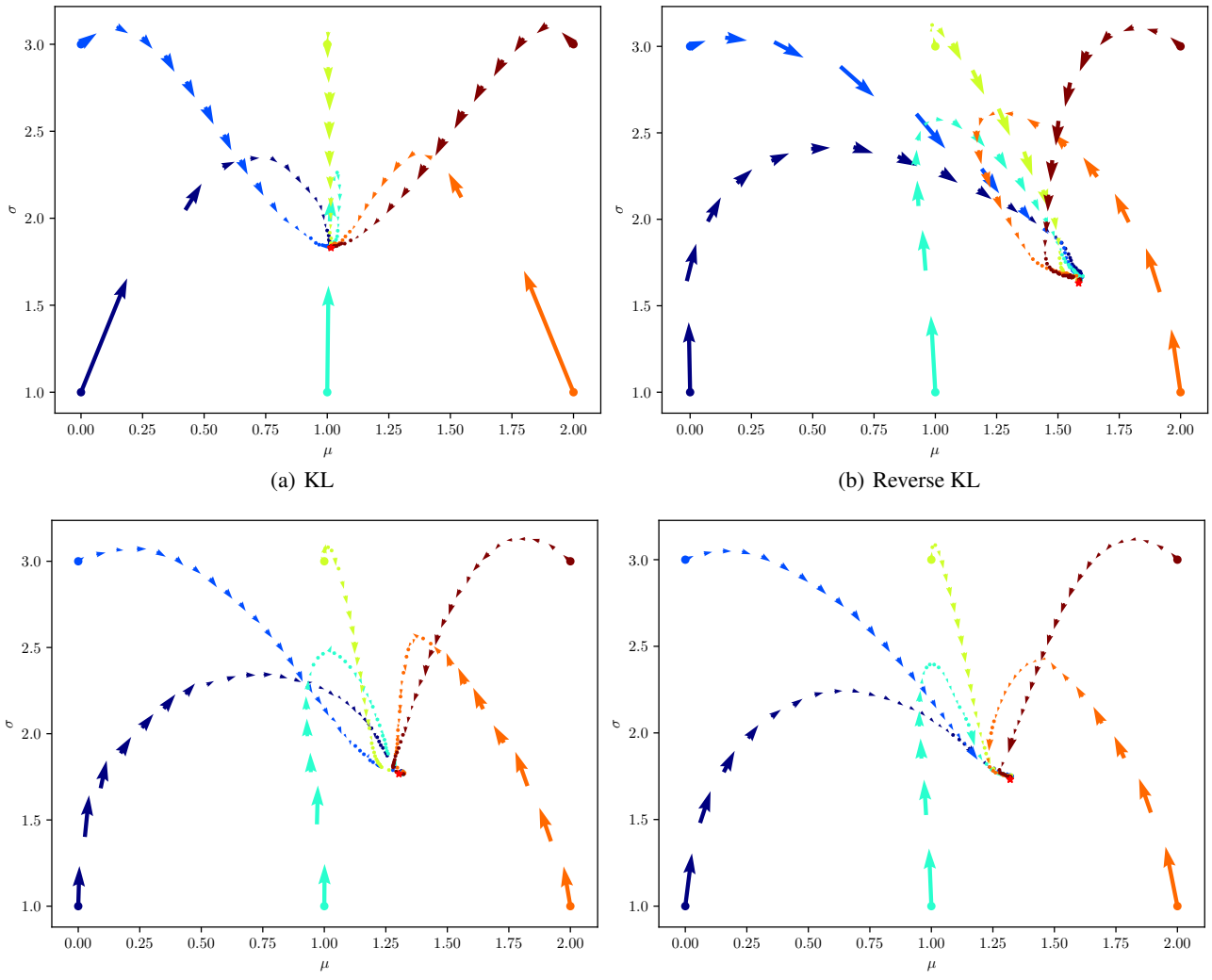


(a) KL (b) Reverse KL

(c) Jensen-Shannon (d) α-Divergence ($\alpha = -1$)

Figure 3. Density ratio estimation results for different f-divergences. The orange solid line represents the ground truth log-scale density ratio ($\log(p(x)/q_\theta(x))$) under a certain divergence; The red dashed line represents the estimated log-scale density ratio ($H_\omega(x) + E_\theta(x)$) learned by f-EBM. Note that the estimated density ratio is accurate in most areas except in the low density regimes (*e.g.* $(-\infty, -2]$ and $[6, +\infty)$) where very few training data comes from this region.

F.3. Optimization Trajectories of Single-Step f-EBM



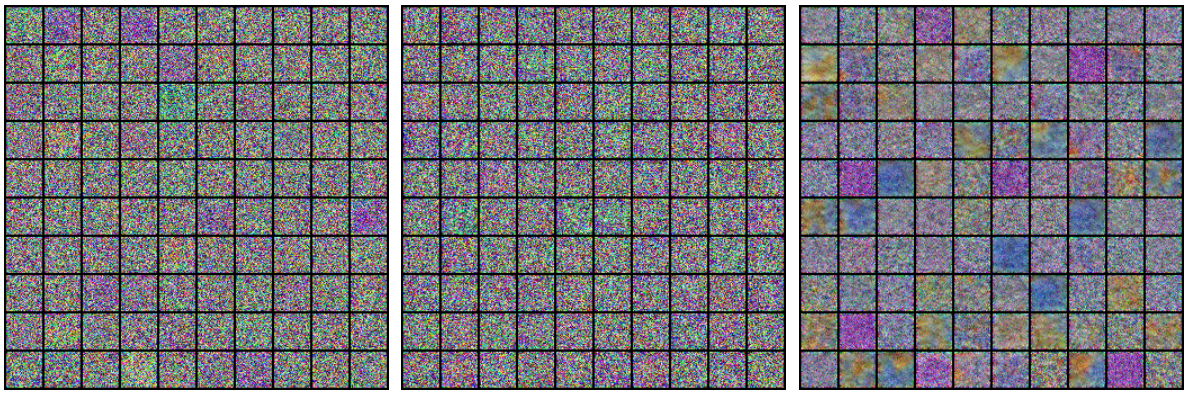
(a) KL (b) Reverse KL

(c) Jensen-Shannon (d) Squared-Hellinger

Figure 4. Convergence of Single-Step f-EBM algorithm for different f-divergences. Optimization trajectories in different colors start from different initializations. The length and direction of the arrows represent the scale and the direction of the gradient at a certain point. The red stars that the trajectories converge to represent the desired optimal solutions under corresponding f-divergences.

G. Additional Experimental Details for Modeling Natural Images

G.1. Samples from The Baseline Method in Section 3.1



(a) Reverse KL (b) Jensen Shannon (c) Squared Hellinger

Figure 5. Uncurated samples from EBMs trained by the baseline approach described in Section 3.1 on CIFAR-10 dataset.

G.2. Uncurated CIFAR-10 Samples from f-EBM

Figure 6. Uncurated CIFAR-10 samples from f-EBM under the guidance of Jensen Shannon.

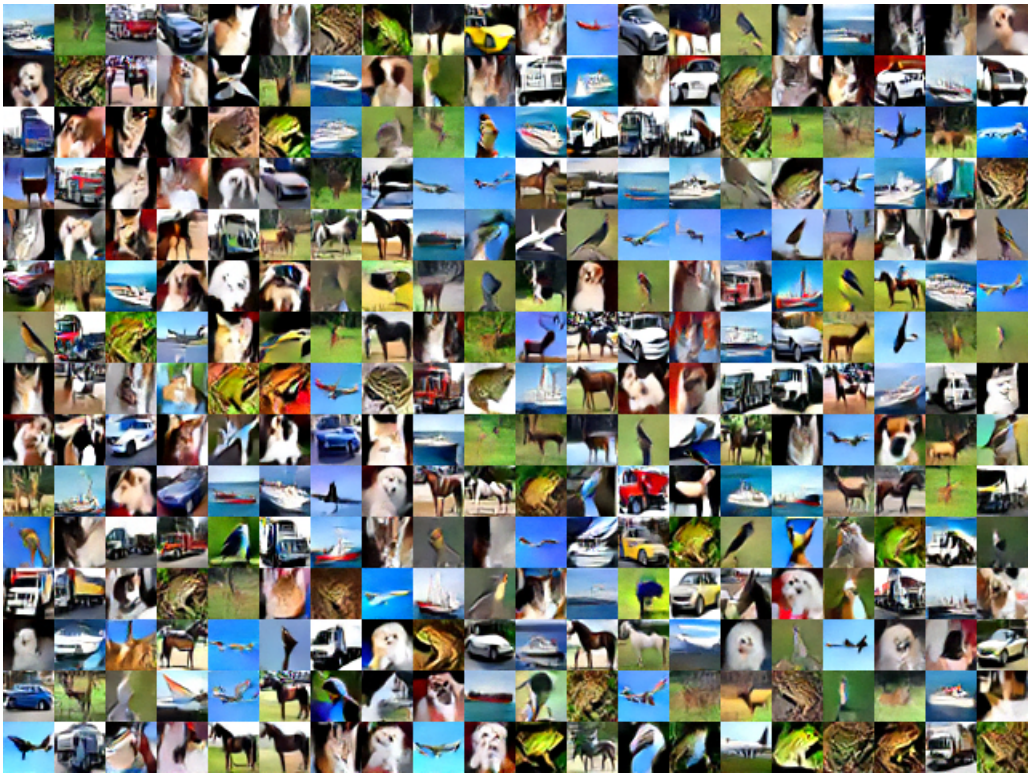


Figure 7. Uncurated CIFAR-10 samples from f-EBM under the guidance of Squared Hellinger.

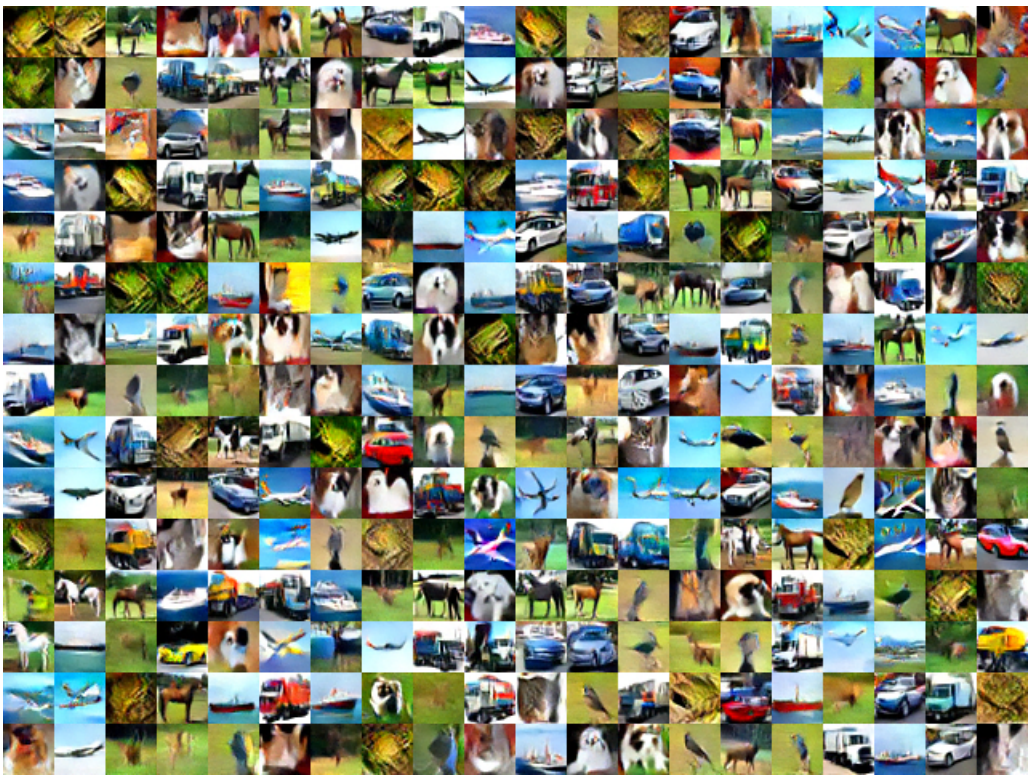
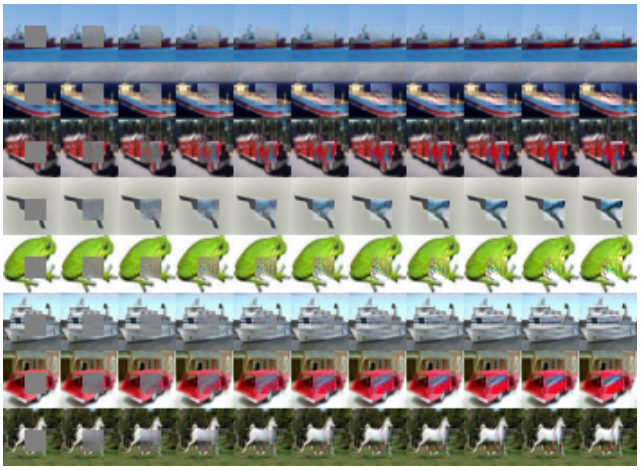


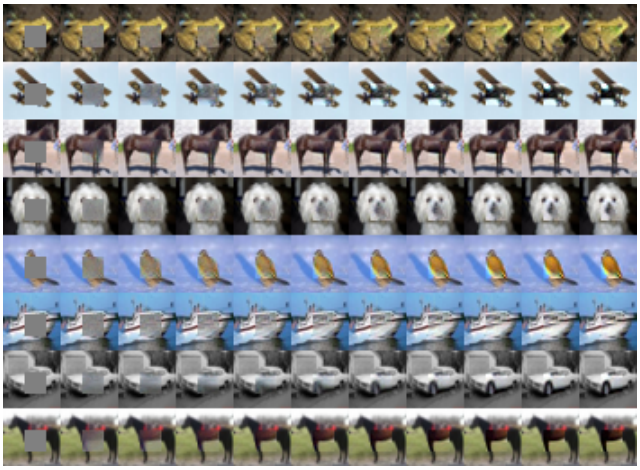
Figure 8. Uncurated CIFAR-10 samples from f-EBM under the guidance of Reverse KL.

G.3. Image Inpainting Results



(a) Reverse KL

(b) Jensen Shannon



(c) Squared Hellinger

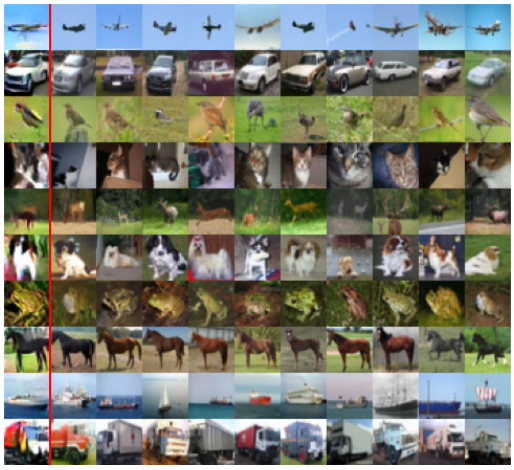
Figure 9. Image inpainting results for f-EBM trained with different f-divergences on CIFAR-10. We use Langevin dynamics sampling to restore the images which are corrupted by empty boxes.

G.4. Image Denoising Results

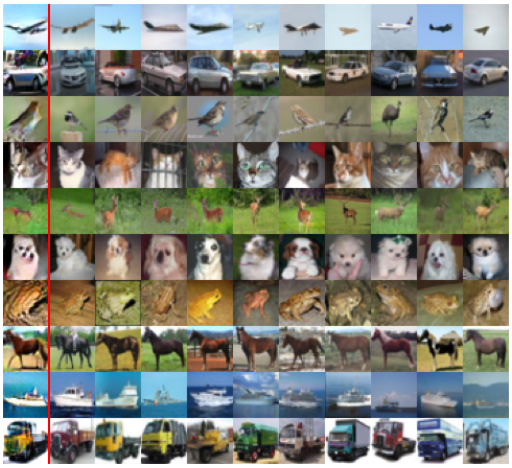
(a) Reverse KL (b) Jensen Shannon (c) Squared Hellinger

Figure 10. Image denoising results for f-EBM trained with different f-divergences on CIFAR-10. We apply 10% "salt and pepper" noise to the images in the test set and use Langevin dynamics sampling to restore the images.

G.5. Nearest Neighbor Images



(a) Reverse KL



(b) Jensen Shannon

(c) Squared Hellinger

Figure 11. Nearest neighbor images according to l_2 distance between images. Different rows are for different classes. In each row, the leftmost image (*i.e.*, on the left of the right vertical line) is generated by f-EBM, and the other images are nearest neighbors of the generated image in the training set.

G.6. Intermediate Samples of Langevin Dynamics on CIFAR-10

(a) Reverse KL

(b) Jensen Shannon

(c) Squared Hellinger

Figure 12. Intermediate samples during Langevin dynamics sampling process for f-EBM.

G.7. Uncurated CelebA Samples from f-EBM

(a) Reverse KL

(b) Jensen Shannon

(c) Squared Hellinger

Figure 13. Uncurated CelebA samples from f-EBM.

G.8. Intermediate Samples of Langevin Dynamics on CelebA

(a) Reverse KL

(b) Jensen Shannon

(c) Squared Hellinger

Figure 14. Intermediate samples during Langevin dynamics sampling process for f-EBM.

G.9. Architectures and Training Hyperparameters

For all the experiments on natural images (conditional generation for CIFAR-10 and unconditional generation for CelebA), we use the residual network architecture (He et al., 2016) in Figure 15 (same as the conditional CIFAR-10 model in (Du & Mordatch, 2019)) to implement both the energy function and the variational function of f-EBM. For the contrastive divergence baseline, we also use the same model architecture for fair comparisons. We note that the model architecture is highly relevant to the model performance and we leave the investigation of better architectures to future works.

3x3 Conv2d, 128
ResBlock Down 128
ResBlock 128
ResBlock Down 256
ResBlock 256
ResBlock Down 256
ResBlock 256
Global Sum Pooling
Dense $\rightarrow$ 1

Figure 15. ResNet architecture for implementing the energy function and the variational function.

For the CelebA dataset, we first center-crop the images to 140×140, then resize them to 32×32. For both CelebA and CIFAR-10, we rescale the images to $[0, 1]$. Following (Du & Mordatch, 2019), we apply spectral normalization and L_2 regularization (on the outputs of the models) with coefficient 1.0 to improve the stability. We use 60 steps Langevin dynamics together with a sample replay buffer of size 10000 to produce samples in the training phase. In each Langevin step, we use a step size of 10.0 and a random noise with standard deviation of 0.005. We use Adam optimizer with $\beta_1 = 0.0, \beta_2 = 0.999$ and learning rate of 3×10^{-4} to optimize the parameters of both the energy function and the variational function. In each training iteration, we use a batch of 128 positive images and negative images. These training hyperparameters are used for both f-EBMs and the contrastive divergence baseline.

G.10. Time Complexity and Convergence Speed

In the experiments we use the same batch size, learning rate and MCMC step number as in CD (Du & Mordatch, 2019). Since we use single-step minimax optimization for both the energy function and the variational function, each iteration needs the same time for MCMC as (Du & Mordatch, 2019) and twice time for stochastic gradient descent parameter updates. Overall, the time complexity per iteration of f-EBM is comparable to that of CD (within a factor of 2). The convergence speed of f-EBM is also similar to CD in terms of number of iterations needed (CD: 75K iterations; f-EBM with Jensen Shannon: 50K iterations; f-EBM with Squared Hellinger: 80K iterations; f-EBM with Reverse KL: 70K iterations; f-EBM with KL: 75K iterations).