Supplementary Material:
STANLEY: Stochastic Gradient Anisotropic Langevin Dynamics for Learning Energy-Based Models

A. Proofs of the Theoretical Results

A.1. Proof of Theorem 1

Theorem. *Assume H1-H3. For any $\theta \in \Theta$, there exists a drift function V_θ, a set $\mathcal{O} \subset \mathcal{Z}$, a constant $0 < \epsilon \leq 1$ such that*

$$\Pi_\theta(z, \mathcal{B}) \geq \epsilon \int_{\mathcal{B}} 1_{\mathcal{X}}(z) dy \ . \tag{14}$$

Moreover there exists $0 < \mu < 1$, $\delta > 0$ and a drift function V, now independent of θ such that for all $z\mathcal{Z}$:

$$\Pi_\theta V(z) \leq \mu V(z) + \delta 1_{\mathcal{O}}(z) \ . \tag{15}$$

Proof. We list the notations used throughout this proof in the following table:

Π_θ	$\triangleq$	Transition kernel of the MCMC defined by (5)
$\mathcal{O}$	$\triangleq$	Subset of $\mathbb{R}^p$ and small set for kernel Π_θ
$B(z, a)$	$\triangleq$	Ball around $z \in \mathcal{Z}$ of radius $a > 0$
$\mathcal{A}_\theta(z)$	$\triangleq$	Acceptance set at state $\in \mathcal{Z}$ such that $\rho_\theta \geq 1$
$\mathcal{A}^*_\theta(z)$	$\triangleq$	Complementary set of $\mathcal{A}_\theta(z)$
$\mathrm{T}_\theta(z', z)$	$\triangleq$	Probability density function of the Gaussian proposal
$\pi_\theta(\cdot)$	$\triangleq$	Stationary/Target distribution under model $\theta \in \Theta$
$\Pi_\theta(z, z')$	$\triangleq$	Transition kernel from state z to state z'
$n_\sigma(z)$	$\triangleq$	Pdf of a centered Normal distribution of standard deviation $\sigma > 0$

The proof of our results are divided into two parts. We first prove the existence of a set noted $\mathcal{O}$ as a small set for our transition kernel Π_θ. Proving a small set is important to show that for any state, the Markov Chain does not stay in the same state, and thus help in proving its irreducibility and aperiodicity.

Then, we will prove the drift condition towards a small set. This condition is crucial to prove the convergence of the chain since it states that the kernels tend to attract elements into that set. finally, uniform ergodicity is established as a consequence of those drift conditions.

(i) Existence of small set: Let $\mathcal{O}$ be a compact subset of the state space $\mathcal{Z}$. We also denote the probability density function (pdf) of the Gaussian proposal of Line 3 as $z \to \mathrm{T}_\theta(z', z)$ for any current state of the chain $z' \in \mathcal{Z}$ and dependent on the EBM model parameter θ. Given STANLEY's MCMC update, at iteration t, the proposal is a Gaussian distribution of mean $z^m_{t-1} + \gamma_t/2 \nabla f_{\theta_t}(z^m_{t-1})$ and covariance $\sqrt{\gamma_t}\mathrm{B}_t$.

We recall the definition of the transition kernel in the case of a Metropolis adjustment and for any model parameter $\theta \in \Theta$ and state $z \in \mathcal{Z}$:

$$\Pi_\theta(z, \mathcal{B}) = \int_{\mathcal{B}} \alpha_\theta(z, y) \mathrm{T}_\theta(z, y) \mathrm{d}y + 1_{\mathcal{B}(z)} \int_{\mathcal{Z}} (1 - \alpha_\theta(z, y)) \mathrm{T}_\theta(z, y) \mathrm{d}y \ , \tag{16}$$

where we have defined the Metropolis ratio between two states $z \in \mathcal{Z}$ and $y \in \mathcal{B}$ as $\alpha_\theta(z, y) = \min(1, \frac{\pi_\theta(z)\mathrm{T}_\theta(z,y)}{\mathrm{T}_\theta(y,z)\pi_\theta(y)})$. Thanks to Assumption H1 and due to the fact that the threshold th leads to a symmetric positive definite covariance matrix with bounded non zero eigenvalues implies that the proposal distribution can be bounded by two zero-mean Gaussian distributions as follows:

$$a n_{\sigma_1}(z - y) \leq \mathrm{T}_\theta(z, y) \leq b n_{\sigma_2}(z - y) \quad \text{for all} \quad \theta \in \Theta \ , \tag{17}$$

where σ_1 and σ_2 are the corresponding standard deviation of the distributions and a and b are some scaling factors.

We denote by ρ_θ the ratio $\frac{\pi_\theta(z)\mathrm{T}_\theta(z,y)}{\mathrm{T}_\theta(y,z)\pi_\theta(y)}$ and given the assumptions H1 and H2, define the quantity

$$\delta = \inf(\rho_\theta(z,y), \theta \in \Theta, \quad z \in \mathcal{O}) > 0 \,. \tag{18}$$

Likewise, the proposal distribution is bounded from below by some quantity noted m. Then,

$$\Pi_\theta(z, \mathcal{B}) \geq \int_{\mathcal{B}\cap\mathcal{X}} \alpha_\theta(z,y)\mathrm{T}_\theta(z,y)\mathrm{d}y \geq \min(1,\delta)m \int_{\mathcal{B}} 1_{\mathcal{X}}(z)\mathrm{d}y \,. \tag{19}$$

Then, given the definition of (18), we can find a compact set $\mathcal{O}$ such that $\Pi_\theta(z,\mathcal{B}) \geq \epsilon$ where $\epsilon = \min(1,\delta)m\mathbf{Z}$ where $\mathbf{Z}$ is the normalizing constant of the pdf $\frac{1}{\mathbf{Z}}\mathbf{1}_{\mathcal{X}}(z)\mathrm{d}y$. The calculations above prove (8), *i.e.,* the existence of a small set for our family of transition kernels $(\Pi_\theta)_{\theta\in\Theta}$.

(ii) Drift condition and ergodicity: We first need to prove the fact that our family of transition kernels $(\Pi_\theta)_{\theta\in\Theta}$ satisfies a drift property.

For a given EBM model parameter $\theta \in \Theta$, we can see in [29] that the drift condition boils down to proving that for the drift function noted V_θ and defined in (6), we have

$$\sup_{z\in\mathcal{Z}} \frac{\Pi_\theta V_\theta(z)}{V_\theta(z)} < \infty \quad \text{and} \quad \lim \sup_{|z|\to\infty} \frac{\Pi_\theta V_\theta(z)}{V_\theta(z)} < 1 \,. \tag{20}$$

Throughout the proof, the model parameter is set to an arbitrary $\theta \in \Theta$. Let denote the acceptation set, *i.e.,* $\rho_\theta \geq 1$ by $\mathcal{A}_\theta(z) := \{y \in \mathcal{Z}, \rho_\theta(z,y) \geq 1\}$ for any state $y \in \mathcal{B}$ and its complementary set $\mathcal{A}^*_\theta(z)$.

STEP (1): Following our definition of the drift function in (6) we obtain:

$$\frac{\Pi_\theta V_\theta(z)}{V_\theta(z)} = \int_{\mathcal{A}_\theta(z)} \mathrm{T}_\theta(z,y)\frac{V_\theta(y)}{V_\theta(z)}\mathrm{d}y + \int_{\mathcal{A}^*_\theta(z)} \frac{\pi_\theta(y)\mathrm{T}_\theta(y,z)}{\pi_\theta(z)\mathrm{T}_\theta(z,y)}\mathrm{T}_\theta(z,y)\frac{V_\theta(y)}{V_\theta(z)}\mathrm{d}y + \int_{\mathcal{A}^*_\theta(z)} (1 - \frac{\pi_\theta(y)\mathrm{T}_\theta(y,z)}{\pi_\theta(z)\mathrm{T}_\theta(z,y)})\mathrm{T}_\theta(z,y)\mathrm{d}y \tag{21}$$

$$\overset{(a)}{\leq} \int_{\mathcal{A}_\theta(z)} \mathrm{T}_\theta(z,y)\frac{\pi_\theta(y)^{-\beta}}{\pi_\theta(z)^{-\beta}}\mathrm{d}y + \int_{\mathcal{A}^*_\theta(z)} \mathrm{T}_\theta(z,y)\frac{\pi_\theta(y)^{1-\beta}}{\pi_\theta(z)^{1-\beta}}\mathrm{d}y + \int_{\mathcal{A}^*_\theta(z)} \mathrm{T}_\theta(z,y)\mathrm{d}y \,, \tag{22}$$

where (a) is due to (6).

According to (17), we thus have that, for any state z in the acceptance set $\mathcal{A}_\theta(z)$:

$$\int_{\mathcal{A}_\theta(z)} \mathrm{T}_\theta(z,y)\frac{\pi_\theta(y)^{-\beta}}{\pi_\theta(z)^{-\beta}}\mathrm{d}y \leq b\int_{\mathcal{A}_\theta(z)} n_{\sigma_2}(y-z)\mathrm{d}y \,. \tag{23}$$

For any state z in the complementary set of the acceptance set, noted $\mathcal{A}^*_\theta(z)$, we also have the following:

$$\int_{\mathcal{A}^*_\theta(z)} \mathrm{T}_\theta(z,y)\frac{\pi_\theta(y)^{1-\beta}}{\pi_\theta(z)^{1-\beta}}\mathrm{d}y \leq \int_{\mathcal{A}^*_\theta(z)} \mathrm{T}_\theta(z,y)^{1-\beta}\mathrm{T}_\theta(y,z)^{\beta}\mathrm{d}y \leq b\int_{\mathcal{A}^*_\theta(z)} n_{\sigma_2}(z-y)\mathrm{d}y \,. \tag{24}$$

While we can define the level set of the stationary distribution π_θ as $\mathcal{L}_{\pi_\theta(y)} = \{z \in \mathcal{Z}, \pi_\theta(z) = \pi_\theta(y)\}$ for some state $y \in \mathcal{B}$, a neighborhood of that level set is defined as $\mathcal{L}_{\pi_\theta(y)}(p) = \{z \in \mathcal{L}_{\pi_\theta(y)}, z + t\frac{z}{|z|}, |t| \leq p\}$. H1 ensures the existence of a radial r such that for all $z \in \mathcal{Z}, |z| \geq r$, then $0 \in \mathcal{L}_{\pi_\theta(y)}$ with $\pi_\theta(z) > \pi_\theta(y)$. Since the function $y \to n_{\sigma_2}(y-z)$ is smooth, it is known that there exists a constant $a > 0$ such that for $\epsilon > 0$, we have that

$$\int_{B(z,a)} n_{\sigma_2}(y-z)\mathrm{d}y \geq 1-\epsilon \quad \text{and} \quad \int_{B(z,a)\cap\mathcal{L}_{\pi_\theta(y)}(p)} n_{\sigma_2}(y-z)\mathrm{d}y \leq \epsilon \,, \tag{25}$$

for some p small enough and where $B(z,a)$ denotes the ball around $z \in \mathcal{Z}$ of radius a. Then combining (23) and (25) we have that:

$$\int_{\mathcal{A}_\theta(z)\cap B(z,a)\cap\mathcal{L}_{\pi_\theta(y)}(p)} \mathrm{T}_\theta(z,y)\frac{\pi_\theta(y)^{-\beta}}{\pi_\theta(z)^{-\beta}}\mathrm{d}y \leq b\epsilon \,. \tag{26}$$

Conversely, we can define the set $\mathcal{A} = \mathcal{A}_\theta(z) \cap B(z,a) \cap \mathcal{L}^+$ where $u \in \mathcal{L}^+$ if $u \in \mathcal{L}_{\pi_\theta(y)}(p)$ and $\phi_\theta(u) > \pi_\theta(p)$. Then using the second part of H1, there exists a radius $r' > r + a$, such that for $z \in \mathcal{Z}$ with $|z| \geq r'$ we have

$$\int_{\mathcal{A}} (\frac{\pi_\theta(y)}{\pi_\theta(z)})^{1-\beta} \mathrm{T}_\theta(y,z)\mathrm{d}y \leq \mathsf{d}(p,r')^{1-\beta} b \int_{\mathcal{A}_\theta(z)} n_{\sigma_2}(y-z)\mathrm{d}y \leq b\mathsf{d}(p,r')^{1-\beta} , \tag{27}$$

where $\mathsf{d}(p,r') = \sup_{|z|>r'} \frac{\pi_\theta(z+p\frac{z}{|z|})}{\pi_\theta(z)}$. Note that H1 implies that $\mathsf{d}(p,r') \to 0$ when $r' \to \infty$. Likewise with $\mathcal{A} = \mathcal{A}_\theta(z) \cap B(z,a) \cap \mathcal{L}^-$ we have

$$\int_{\mathcal{A}} (\frac{\pi_\theta(y)}{\pi_\theta(z)})^{-\beta} \mathrm{T}_\theta(z,y)\mathrm{d}y \leq b\mathsf{d}(p,r')^{\beta} . \tag{28}$$

Same arguments can be obtained for the second term of (21), *i.e.,* $\mathrm{T}_\theta(z,y)\frac{\pi_\theta(y)^{1-\beta}}{\pi_\theta(z)^{1-\beta}}$ and we obtain, plugging the above in (21) that:

$$\lim\sup_{|z|\to\infty} \frac{\Pi_\theta V_\theta(z)}{V_\theta(z)} \leq \lim\sup_{|z|\to\infty} \int_{\mathcal{A}^*_\theta(z)} \mathrm{T}_\theta(z,y)\mathrm{d}y . \tag{29}$$

Since $\mathcal{A}^*_\theta(z)$ is the complementary set of $\mathcal{A}_\theta(z)$, the above inequality yields

$$\lim\sup_{|z|\to\infty} \frac{\Pi_\theta V_\theta(z)}{V_\theta(z)} \leq 1 - \lim\inf_{|z|\to\infty} \int_{\mathcal{A}_\theta(z)} \mathrm{T}_\theta(z,y)\mathrm{d}y . \tag{30}$$

STEP (2): The final step of our proof consists in proving that $1 - \lim\inf_{|z|\to\infty} \int_{\mathcal{A}_\theta(z)} \mathrm{T}_\theta(z,y)\mathrm{d}y \leq 1 - c$ where c is a constant, independent of all the other quantities.

Given that the proposal distribution is a Gaussian and using assumption H1 we have the existence of a constant c_a depending on a as defined above (the radius of the ball $B(z,a)$) such that

$$\frac{\pi_\theta(z)}{\pi_\theta(z - \ell\frac{z}{|z|})} \leq c_a \leq \inf_{y \in B(z,a)} \frac{\mathrm{T}_\theta(y,z)}{\mathrm{T}_\theta(z,y)} \quad \text{for any } z \in \mathcal{Z}, |z| \geq r^* . \tag{31}$$

Then for any $|z| \geq r^*$, we obtain that $z - \ell\frac{z}{|z|} \in \mathcal{A}_\theta(z)$. A particular subset of $\mathcal{A}_\theta(z)$ used throughout the rest of the proof is the cone defined as

$$\mathcal{P}(z) := \{z - \ell\frac{z}{|z|} - \kappa\nu, \text{ with } \quad i < a - \ell, \nu \in \{\nu \in \mathbb{R}^d, \|\nu\| < 1\}, |\nu - \frac{z - \ell\frac{z}{|z|}}{|z - \ell\frac{z}{|z|}|} \leq \frac{\epsilon}{2}\} . \tag{32}$$

Using Lemma 1, we have that $\mathcal{P}(z) \subset \mathcal{A}_\theta(z)$

Then,

$$\int_{\mathcal{A}_\theta(z)} \mathrm{T}_\theta(z,y)\mathrm{d}y \overset{(a)}{\geq} \int_{\mathcal{A}_\theta(z)} a n_{\sigma_1}(y-z)\mathrm{d}y \overset{(b)}{\geq} a \int_{\mathcal{P}(z)} n_{\sigma_1}(y-z)\mathrm{d}y , \tag{33}$$

where we have used (17) in (a) and applied Lemma 1 in (b).

If we define the translation of vector $z \in \mathcal{Z}$ by the operator $\mathcal{I} \subset \mathbb{R}^d \to T_z(\mathcal{I})$, then

$$\int_{\mathcal{A}_\theta(z)} \mathrm{T}_\theta(z,y)\mathrm{d}y \geq a \int_{\mathcal{P}(z)} n_{\sigma_1}(y-z)\mathrm{d}y = \int_{T_z(\mathcal{P}(z))} n_{\sigma_1}(y-z)\mathrm{d}y . \tag{34}$$

Recalling the objective of STEP (2) that is to find a constant c such that $1 - \lim\inf_{|z|\to\infty} \int_{\mathcal{A}_\theta(z)} \mathrm{T}_\theta(z,y)\mathrm{d}y \leq 1 - c$, we see from (34) that since the set $\mathcal{P}(z)$ does not depend on the EBM model parameter θ and that once translated by z the resulting set $T_z(\mathcal{P}(z))$ is independent of z (but depends on ℓ, see definition (32), then the integral $\int_{T_z(\mathcal{P}(z))} n_{\sigma_1}(y-z)\mathrm{d}y$ in (34) is independent of z thus concluding on the existence of the constant c such that

$$\lim\sup_{|z|\to\infty} \frac{\Pi_\theta V_\theta(z)}{V_\theta(z)} \leq 1 - c .$$

Thus proving the second part of (20) which is the main drift condition we ought to demonstrate. The first part of (20) can be proved by observing that $\frac{\Pi_\theta V_\theta(z)}{V_\theta(z)}$ is smooth on $\mathcal{Z}$ according to H2 and by construction of the transition kernel. Smoothness implies boundedness on the compact $\mathcal{Z}$.

STEP (3): We now use the main proven equations in (20) to derive the second result (9) of Theorem 1.

We will begin by showing a similar inequality for the drift function V_θ, thus not having uniformity, as an intermediary step. The Drift property is a consequence of STEP (2) and (34) shown above. Thus, there exists $0 < \bar{\mu} < 1$, $\bar{\delta} > 0$ such that for all $z\mathcal{Z}$:

$$\Pi_\theta V_\theta(z) \leq \bar{\mu} V_\theta(z) + \bar{\delta} 1_{\mathcal{O}}(z) \,, \tag{35}$$

where V_θ is defined by (6).

Using the two functions defined in (7), we define for $z \in \mathcal{Z}$, the V function independent of θ as follows:

$$V(z) = V_1(z)^\alpha V_2(z)^{2\alpha} \,, \tag{36}$$

where $0 < \alpha < \min(\frac{1}{2\beta}, \frac{a_0}{3})$, a_0 is defined in H3 and β is defined in (6). Thus for $\theta \in \Theta$, $z \in \mathcal{Z}$ and $\epsilon > 0$:

$$\begin{aligned}
\Pi_\theta V(z) &= \int_{\mathcal{Z}} \Pi_\theta(z,y) V_1(y)^\alpha V_2(y)^{2\alpha} \mathrm{d}y \\
&\overset{(a)}{\leq} \frac{1}{2} \int_{\mathcal{Z}} \Pi_\theta(z,y) (\frac{1}{\epsilon^2} V_1(y)^{2\alpha} + \epsilon^2 V_2(y)^{4\alpha}) \mathrm{d}y \,, \\
&\overset{(b)}{\leq} \frac{1}{2\epsilon^2} \int_{\mathcal{Z}} \Pi_\theta(z,y) V_\theta(y)^{2\alpha} + \frac{\epsilon^2}{2} \int_{\mathcal{Z}} \Pi_\theta(z,y) V_2(y)^{4\alpha} \mathrm{d}y \,,
\end{aligned} \tag{37}$$

where we have used the Young's inequality in (a) and the definition of V_1, see (7), in (b). Then plugging (35) in (37), we have

$$\Pi_\theta V(z) \leq \frac{1}{2\epsilon^2} (\bar{\mu} V_\theta(z)^{2\alpha} + \bar{\delta} 1_{\mathcal{O}}(z)) + \frac{\epsilon^2}{2} \int_{\mathcal{Z}} \Pi_\theta(z,y) V_2(y)^{4\alpha} \mathrm{d}y \,, \tag{38}$$

$$\leq \frac{\bar{\mu}}{2\epsilon^2} V(z) + \frac{\bar{\delta}}{2\epsilon^2} 1_{\mathcal{O}}(z) + \frac{\epsilon^2}{2} \int_{\mathcal{Z}} \Pi_\theta(z,y) V_2(y)^{4\alpha} \mathrm{d}y \,, \tag{39}$$

$$\leq \frac{\bar{\mu}}{2\epsilon^2} V(z) + \frac{\bar{\delta}}{2\epsilon^2} 1_{\mathcal{O}}(z) + \frac{\epsilon^2}{2} \sup_{\theta \in \Theta, z \in \mathcal{Z}} \int_{\mathcal{Z}} \Pi_\theta(z,y) V_2(y)^{4\alpha} \mathrm{d}y \,, \tag{40}$$

$$\leq \frac{\bar{\mu}}{2\epsilon^2} V(z) + \frac{\bar{\delta}}{2\epsilon^2} 1_{\mathcal{O}}(z) + \frac{\epsilon^2}{1+\bar{\mu}} V(z) \,, \tag{41}$$

$$\leq \left(\frac{\bar{\mu}}{2\epsilon^2} + \frac{\epsilon^2}{1+\bar{\mu}} \right) V(z) + \frac{\bar{\delta}}{2\epsilon^2} 1_{\mathcal{O}}(z) \,, \tag{42}$$

where we have used (36) and the assumption H3 in the last inequality, ensuring the existence of such exponent α.

Setting $\epsilon := \sqrt{\frac{\bar{\mu}(1+\bar{\mu})}{2}}$, $\mu := \sqrt{\frac{2\bar{\mu}}{1+\bar{\mu}}}$ and $\delta := \frac{\bar{\delta}}{2\epsilon^2}$ proves the uniform ergodicity in (9) and concludes the proof of Theorem 1. □

A.2. Proof of Lemma 1

Lemma. *Define* $\mathcal{P}(z) := \{z - \ell \frac{z}{|z|} - \kappa\nu$, *with* $\kappa < a - \ell, \nu \in \{\nu \in \mathbb{R}^d, \|\nu\| < 1\}, |\nu - \frac{z - \ell \frac{z}{|z|}}{|z - \ell \frac{z}{|z|}|}| \leq \frac{\epsilon}{2}\}$ *and* $\mathcal{A}_\theta(z) := \{y \in \mathcal{Z}, \rho_\theta(z,y) \geq 1\}$. *Then for* $z \in \mathcal{Z}$, $\mathcal{P}(z) \subset \mathcal{A}_\theta(z)$.

Proof. In order to show the inclusion of the set $\mathcal{P}(z)$ in $\mathcal{A}_\theta(z)$ we start by selecting the quantity $y = z - \ell \frac{z}{|z|} - \kappa\nu$ for $z \in \mathcal{Z}$ and $\kappa < a - \ell$ where a is the radius of the ball used in (25) such that $y \in \mathcal{P}(z)$. We will now show that $y \in \mathcal{A}_\theta(z)$.

By the generalization of Rolle's theorem applied on the stationary distribution π_θ, we guarantee the existence of some κ^* such that:

$$\nabla \pi_\theta(z - \ell \frac{z}{|z|} - \kappa^* \nu) = \frac{\pi_\theta(y) - \pi_\theta(z - \ell \frac{z}{|z|})}{y - (z - \ell \frac{z}{|z|})} \tag{43}$$

$$= -\frac{\pi_\theta(y) - \pi_\theta(z - \ell \frac{z}{|z|})}{\kappa\nu} \,. \tag{44}$$

Expanding $\nabla\pi_\theta(z - \ell\frac{z}{|z|} - \kappa^*\nu)$ yields:

$$\pi_\theta(y) - \pi_\theta(z - \ell\frac{z}{|z|}) = -\kappa\nu\frac{z - \ell\frac{z}{|z|} - \kappa^*\nu}{|z - \ell\frac{z}{|z|} - \kappa^*\nu|}|\nabla\pi_\theta(z - \ell\frac{z}{|z|} - \kappa^*\nu)| \,. \tag{45}$$

Yet, under H1, there exists ϵ such that

$$\frac{\nabla f_\theta(z)}{|\nabla f_\theta(z)|}\frac{z}{|z|} \leq -\epsilon \,,$$

and for any $y \in \mathcal{P}(z)$ we note that $\frac{y}{|y|} - \frac{z}{|z|}| \leq \frac{\epsilon}{2}$, by construction of the set. Thus,

$$\frac{\nabla f_\theta(y)}{|\nabla f_\theta(y)|}\nu = \frac{\nabla f_\theta(y)}{|\nabla f_\theta(y)|}(\nu - \frac{z - \ell\frac{z}{|z|}}{|z - \ell\frac{z}{|z|}|}) + \frac{\nabla f_\theta(y)}{|\nabla f_\theta(y)|}(\nu - \frac{z - \ell\frac{z}{|z|}}{|z - \ell\frac{z}{|z|}|} - \frac{y}{|y|}) + \frac{\nabla f_\theta(y)}{|\nabla f_\theta(y)|}\frac{y}{|y|} \leq 0 \,, \tag{46}$$

where ν is used in the definition of $\mathcal{P}(z)$. Also note that $\frac{\nabla f_\theta(y)}{|\nabla f_\theta(y)|}\nu$ denotes the vector multiplication between the normalized gradient and ν.

Then plugging (46) into (45) leads to $\pi_\theta(y) - \pi_\theta(z - \ell\frac{z}{|z|}) \geq 0$. Then $y \in \mathcal{P}(z)$ implies, using (31), that $\pi_\theta(y) \geq \pi_\theta(z - \ell\frac{z}{|z|}) \geq \frac{1}{c_a}\pi_\theta(z)$. Finally $y \in \mathcal{P}(z)$ implies that $y \in \mathcal{A}_\theta(z)$, concluding the proof of Lemma 1.

□