The forward–backward envelope for sampling with the overdamped Langevin algorithm

Armin Eftekhari · Luis Vargas · Konstantinos C. Zygalakis

Received: 24 January 2022 / Accepted: 2 May 2023 / Published online: 31 May 2023
© The Author(s) 2023

Abstract
In this paper, we analyse a proximal method based on the idea of forward–backward splitting for sampling from distributions with densities that are not necessarily smooth. In particular, we study the non-asymptotic properties of the Euler–Maruyama discretization of the Langevin equation, where the forward–backward envelope is used to deal with the non-smooth part of the dynamics. An advantage of this envelope, when compared to widely-used Moreau–Yoshida one and the MYULA algorithm, is that it maintains the MAP estimator of the original non-smooth distribution. We also study a number of numerical experiments that support our theoretical findings.

Keywords Markov chain Monte Carlo · Convex optimization · Langevin equation

1 Introduction
The problem of calculating expectations with respect to a probability distribution $p$ in $\mathbb{R}^d$ is ubiquitous throughout applied mathematics, statistics, molecular dynamics, statistical physics and other fields.

In practice, often the dimension $d$ is large, which renders deterministic techniques, such as quadrature methods, computationally intractable. In contrast, probabilistic methods do not suffer from this curse of dimensionality and are often the method of choice when the dimension $d$ is large.

In particular, Markov chain Monte Carlo (MCMC) methods are based on the construction of a Markov chain in $\mathbb{R}^m$ with $m \geq d$, for which the invariant distribution (or its suitable marginal) coincides with the target distribution $p$ (Brooks et al. 2011).

Often, such Markov chains are based on the discretization of stochastic differential equations (SDEs). One such SDE, which is also the focus of this paper, is the (overdamped) Langevin equation

$$dX_t = -\nabla f(X_t) \, dt + \sqrt{2} \, dW_t,$$  

where $\{W_t\}$ is the standard $d$-dimensional Brownian motion and $\nabla f$ denotes the gradient of a continuously-differentiable function $f : \mathbb{R}^d \to \mathbb{R}$. Under mild assumptions on $f$, the dynamics of (1) are ergodic with respect to the distribution $p \propto e^{-f}$. In particular, $p$ is the invariant distribution of (1); see Milstein and Tretyakov (2007).

The discretization of (1), however, requires special care because the resulting discrete Markov chain might not be ergodic anymore (Roberts and Tweedie 1996). In addition, even if ergodic, the resulting discrete Markov chain often has a different invariant distribution than $p$, known as the numerical invariant distribution $\hat{p}$. The study of the asymptotic error between the numerical invariant distribution $\hat{p}$ and the target distribution $p$ has received considerable attention recently (Mattingly et al. 2010; Abdulle et al. 2014). In particular, Mattingly et al. (2010) investigated the effect of discretization on the convergence of the ergodic averages, and Abdulle et al. (2014) presented general order conditions to ensure that...
the numerical invariant distribution accurately approximates the target distribution.

Another active line of research quantifies the nonasymptotic error between the numerical invariant distribution $\hat{p}$ and the target distribution $p$. In particular, when $p$ is a smooth and strongly log-concave distribution, Dalalyan (2017) established non-asymptotic bounds in total variation distance for the Euler–Maruyama discretization of (1), commonly known as the unadjusted Langevin algorithm (ULA). These results have also been extended to the Wasserstein distance $W_2$ and the KL divergence in Durmus and Moulines (2017), Dalalyan (2017), Durmus and Moulines (2019), Dalalyan and Karagulyan (2019), Durmus et al. (2019), to name a few. Typically, these works study the number of iterations that the numerical integrator would require to achieve a desired accuracy, when applied to a target distribution $p$ with a known condition number.

In fact, the above strong log-concavity of $p$ can be substantially relaxed. In particular, using a variant of the reflection coupling, the recent work (Eberle and Majka 2019) derives non-asymptotic bounds for the ULA in the Wasserstein distance $W_1$, when $p$ is strictly log-concave outside of a ball in $\mathbb{R}^d$. Similar results for the Wasserstein distance $W_2$ have been obtained in Majka et al. (2020).

Within the class of log-concave distributions, a significant challenge for the Langevin diffusion in (1) arises when the target distribution $p$ is not smooth and/or has a compact (convex) support in $\mathbb{R}^d$. One approach to address this challenge is to replace the non-smooth distribution $p$ with a smooth proxy obtained via the so-called Moreau–Yoshida (MY) envelope. This new smooth density remains log-concave and thus amenable to the non-asymptotic results discussed earlier. When the support of $p$ is also compact, proximal Monte Carlo methods have been explored in Brosse et al. (2017), Pereyra (2016), Durmus et al. (2018). It is also worth noting that (Bubeck et al. 2018) pursued a different approach for sampling from compactly-supported densities that does not involve the MY envelope.

A potential drawback of the above approach is that the MY envelope often does not maintain the maximum a posteriori (MAP) estimator. That is, the above approach alters the location at which the new (smooth) density reaches its maximum. This is a well-known issue in the context of (non-smooth) convex optimization and is often resolved by appealing to the proximal gradient method. The latter can be understood as the Euler discretization of the so-called forward–backward (FB) envelope (Stella et al. 2017).

1.1 Contributions

This work explores and analyzes the use of the FB envelope for sampling from non-smooth and compactly-supported log-concave distributions. In analogy with the Langevin proximal Monte Carlo, we replace the non-smooth density with a smooth proxy obtained via the FB envelope. In particular, this proxy is strongly log-concave over long distances.

Crucially, the new proxy also maintains the MAP estimator under certain assumptions. However, this improvement comes at the cost of requiring additional smoothness for the smooth part of the density. Lastly, the strong convexity of the new proxy over long distances allows us to utilise the work of Eberle and Majka (2019) to obtain non-asymptotic guarantees for our method in the Wasserstein distance $W_1$.

In addition to investigating the use of the FB envelope in sampling, this work has the following contributions:

- It introduces a general theoretical framework for sampling from non-smooth densities by introducing the notion of admissible envelopes. MY and FB envelopes are both instances of admissible envelopes.
- It proposes a new Langevin algorithm to sample from non-smooth densities, dubbed EULA. EULA generalizes MYULA in the sense that it can work with any admissible envelope, e.g., MY or FB. EULA can also handle a family of increasingly more accurate envelopes rather than using a fixed envelope.

1.2 Organisation

The rest of the paper is organised as follows. Section 2 formalizes the problem of sampling from a non-smooth and compactly-supported log-concave distribution. As a proxy for this non-smooth distribution, its (smooth) Moreau–Yosida (MY) envelope is reviewed in Sect. 3. This section also explains the main limitation of MY envelope, i.e., its inaccurate MAP estimation. In Sect. 4, we introduce the forward–backward (FB) envelope which overcomes the key shortcoming of the MY envelope.

Section 5 introduces and analyses EULA, an extension of the popular ULA for sampling from a non-smooth distribution. EULA can be adapted to various envelopes. In particular, MYULA from Brosse et al. (2017) is a special case of EULA for the MY envelope. Section 6 proves the iteration complexity of EULA and Sect. 7 presents a few numerical examples to support the theory developed here.

2 Statement of the problem

Consider a compact convex set $K \subset \mathbb{R}^d$. For a pair of functions $\overline{f} : \mathbb{R}^d \to \mathbb{R}$ and $\overline{g} : \mathbb{R}^d \to \mathbb{R}$, our objective in this work is to sample from the probability distribution

$$p(x) := \begin{cases} \frac{e^{-\overline{f}(x) - \overline{g}(x)}}{\int_K e^{-\overline{f}(z) - \overline{g}(z)} \, dz} & x \in K, \\ 0 & x \notin K, \end{cases}$$

(2)
whenever the ratio above is well-defined. In order to sample from $p$, we only have access to the gradient of $\tilde{F}$ and the proximal operator for $\bar{g}$, to be defined later. Our assumptions on $K, \tilde{F}, \bar{g}$ are detailed below.

**Assumption 1** We make the following assumptions:

1. For radii $R \geq r > 0$, assume that $K \subset \mathbb{R}^d$ is a compact convex body that satisfies $B(0, r) \subset K \subset B(0, R)$. Here, $B(0, r)$ is the Euclidean ball of radius $r$ centered at the origin.
2. Assume also that $\tilde{F} : \mathbb{R}^d \to \mathbb{R}$ is a convex function that is three-times continuously differentiable.
3. Assume lastly that $\bar{g} : \mathbb{R}^d \to (-\infty, \infty]$ is a proper closed convex function. Moreover, we assume that $\bar{g}$ is continuous.\(^1\)

A few important remarks about Assumption 1 are in order. First, in the special case when $\tilde{F}$ is a convex quadratic (Luu et al. 2021), the assumption of thrice-differentiability above is trivially met and some of the developments below are simplified. However, our more general setup here necessitates the thrice-differentiability above and leads below to more involved technical derivations.

Second, for the sake of mathematical convenience, instead of the two functions $\tilde{F}, \bar{g}$, we will work throughout this manuscript with two new functions $f, g$, without any loss of generality.

More specifically, instead of $\tilde{F}$, we consider a convex function $f$ that coincides with $\tilde{F}$ on the set $K$, has a compact support and a continuously differentiable Hessian. For this function $f$, the compactness of $K$ and the smoothness of $\tilde{F}$ in Assumption 1 together imply that $f, \nabla f, \nabla^2 f$ are all Lipschitz-continuous functions. To summarize, for the function $f$ described above, there exist nonnegative constants $\lambda_0, \lambda_1, \lambda_2, \lambda_3$ such that

$$f(x) = 0, \quad \text{if } \|x\|_2 \geq \lambda_0, \quad (3a)$$

$$|f(x) - f(y)| \leq \lambda_1 \|x - y\|_2, \quad (3b)$$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \lambda_2 \|x - y\|_2, \quad (3c)$$

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq \lambda_3 \|x - y\|_2, \quad (3d)$$

for every $x, y \in \mathbb{R}^d$. In particular, the compactness of $K$ and the differentiability of $\tilde{F}$ in Assumption 1 imply that $f$ is finite, i.e., $\max_{x \in K} |f(x)| < \infty$.

Instead of $\bar{g}$, we consider the proper closed convex function $g$, defined as

$$g := \bar{g} + 1_K, \quad (4)$$

where $1_K$ is the indicator function for the set $K$. That is, $1_K(x) = 0$ if $x \in K$ and $1_K(x) = \infty$ if $x \notin K$. The compactness of $K$ and the continuity of $\bar{g}$ in Assumption 1 together imply that $g$ is finite, when its domain is limited to the set $K$. Outside of the set $K$, $g$ is infinite. To summarize, the new function $g$ is lower semi-continuous and also satisfies

$$\max_{x \in K} |g(x)| < \infty, \quad g(x) = \infty, \quad \text{if } x \notin K. \quad (5)$$

Using the new functions $f$ and $g$, we rewrite the definition of $p$ in (2) as

$$p(x) := \frac{e^{-F(x)}}{\int_{\mathbb{R}^d} e^{-F(z)} \, dz}, \quad (6a)$$

$$F(x) := \tilde{F}(x) + \bar{g}(x) + 1_K(x) = f(x) + g(x). \quad (6b)$$

Above, $F$ is often referred to as the potential associated with $p$. The last identity above holds by construction. Indeed, on the set $K$, the functions $f$ and $\tilde{F}$ coincide. Likewise, on the set $K$, the functions $g$ and $\bar{g}$ coincide. Outside of the set $K$, however, both sides of the last equality above are infinite. In particular, $F$ is finite on $K$ and infinite outside of this set.

In view of (6b), instead of $\tilde{F}$ and $\bar{g}$, we can work with $f$ and $g$ without any change to the target distribution $p$. This will be more convenient for the proofs. As a side note, Assumption 1 implies that $p$ is finite; see after (3) and (5)–(6). The integral in the denominator in (6a) is also finite by Assumption 1. When there is no confusion, we will overload our notation and use $p$ to also denote the probability measure associated with the law $p$.

Since $g$ is not differentiable, $F$ in (6b) is itself non-differentiable. In turn, this means that one cannot provably use gradient based algorithms (such as the ULA) to sample from $p \propto \exp(-F)$ (Dalalyan and Karagulyan 2019). One way to deal with this issue is to replace $F$ with a smooth function $F_\gamma$, which we will refer to as an envelope. To this envelope we can then apply the ULA. It is reasonable to require this envelope $F_\gamma$ to fulfill the following admissibility assumptions.

**Definition 1** (Admissible Envelopes) For $\gamma^0 > 0$, the functions $\{ F_\gamma : \gamma \in (0, \gamma^0) \}$ are admissible envelopes of $F$ if

1. There exists a function $F^0 : \mathbb{R}^d \to [-\infty, \infty]$ such that $e^{-F^0}$ is integrable, and $F_\gamma$ dominates $F^0$. That is, for every $x \in \mathbb{R}^d$ and $\gamma \in (0, \gamma^0)$, it holds that $F_\gamma(x) \geq F^0(x)$.
2. $F_\gamma : \mathbb{R}^d \to [-\infty, \infty]$ converges pointwise to $F$. That is, $\lim_{\gamma \to 0} F_\gamma(x) = F(x)$ for every $x \in \mathbb{R}^d$.
3. \( F_y \) is \( \lambda_y \)-smooth, i.e., there exists a constant \( \lambda_y \geq 0 \) such that for all \( x, y \in \mathbb{R}^d \)
\[
\| \nabla F_y(x) - \nabla F_y(y) \|_2 \leq \lambda_y \| x - y \|_2, \quad y \in (0, y^0).
\]
If \( \{ F_{y_k} : y \in (0, y^0) \} \) are admissible envelopes of \( F \), we can define the corresponding probability densities: For every \( x \in \mathbb{R}^d \) and every \( \gamma \in (0, y^0) \), we define
\[
p_y(x) := \frac{e^{-F_y(x)}}{\int_{\mathbb{R}^d} e^{-F_y(z)} \, dz}.
\]

**Remark 1** Definition 1.1 implies that \( p_y \) can be normalized and (7) is thus well-defined. This observation and Definition 1.2 together imply, after an application of the dominated convergence theorem, that
\[
\lim_{\gamma \to 0^+} p_y(x) = p(x), \quad x \in \mathbb{R}^d,
\]
where the probability measures \( p \) and \( p_y \) are defined in (6a) and (7), respectively. That is, \( p_y \) converges weakly to \( p \) in the limit of \( \gamma \to 0 \). In other words, we can use \( p_y \) as a proxy for \( p \), provided that \( \gamma \) is sufficiently small. Finally, as we will see shortly, Definition 1.3 will help us establish the convergence of the ULA to an invariant distribution close to \( p_y \), provided that the step size of the ULA is small.

### 3 Moreau–Yosida envelope and its limitation

For \( \gamma > 0 \), let us define
\[
F_y^{\text{MY}}(x) := f(x) + g_y(x),
\]
where
\[
g_y(x) := \min_{z \in \mathbb{R}^d} \left\{ g(z) + \frac{1}{2\gamma} \| x - z \|^2 \right\}
\]
is the Moreau–Yosida (MY) envelope of \( g \). Somewhat inaccurately, we will also refer to \( F_y^{\text{MY}} \) as the MY envelope of \( F \), to distinguish \( F_y^{\text{MY}} \) from its alternative later in this paper. It is well-known that \( g_y \) is \( \gamma^{-1} \)-smooth and that \( g_y \) converges pointwise to \( g \) in the limit of \( \gamma \to 0 \). These facts enable us to establish the admissibility of MY envelopes, as detailed below. All proofs are deferred to the appendices. We note that the result below closely relates to Durmus et al. (2018, Proposition 1). In effect, the result below restates the result in Durmus et al. (2018) in the framework of admissible envelopes.

**Proposition 1** (Admissibility of MY envelopes) Suppose that Assumption 1 is fulfilled. Then \( \{ F_{y}^{\text{MY}} : \gamma > 0 \} \) are admissible envelopes of \( F \) in (6b). In particular, \( \nabla F_y^{\text{MY}} \) is \((\lambda_2 + \gamma^{-1})\)-Lipschitz continuous, and given by the expression
\[
\nabla F_y^{\text{MY}}(x) = \nabla f(x) + \frac{x - p_y(x)}{\gamma},
\]
\[
P_y(x) := \arg \min_{z \in \mathbb{R}^d} \left\{ g(z) + \frac{1}{2\gamma} \| x - z \|^2 \right\}.
\]

Above, \( \lambda_2 \) was defined in (3c), and \( P_y : \mathbb{R}^d \to \mathbb{R}^d \) is the proximal operator associated with the function \( \gamma g \).

**Remark 2** (Connection to Nesterov’s smoothing technique) Alternatively, we can also view the MY envelope through the lens of Nesterov’s smoothing technique (Nesterov 2005). More specifically, if Assumption 1 is fulfilled, one can invoke a standard minimax theorem to verify that
\[
g_y(x) = \max_{z \in \mathbb{R}^d} \left\{ (x, z) - g^*(z) - \frac{\gamma}{2} \| z \|^2 \right\}, \quad x \in \mathbb{R}^d,
\]
where \( g^* \) is the Fenchel conjugate of \( g \). This interpretation is central to Nesterov’s technique for minimizing the non-smooth function \( F \) in (6a).

In view of the admissibility of \( F_y^{\text{MY}} \) by Proposition 1, we can apply the ULA to the new (smooth) potential \( F_y^{\text{MY}} \) instead of the nonsmooth \( F \). In particular, if \( \gamma \) is sufficiently small, then \( p_y \propto \exp(-F_y^{\text{MY}}) \) would be close to the target distribution \( p \) by Remark 1. This technique is known as the MYULA (Brosse et al. 2017).

However, a limitation of the MY envelope is that the minimizers of \( F_y^{\text{MY}} \) are not necessarily the same as the minimizers of \( F \). In turn, the MAP estimator of \( p_y^{\text{MY}} \), denoted by \( \chi_y^{\text{MY}} \), might not coincide with the MAP estimator of \( p \), except in the limit of \( \gamma \to 0 \). That is,
\[
\lim_{\gamma \to 0^+} F_y^{\text{MY}}(\chi_y^{\text{MY}}) = \min_F F(x).
\]
This observation is particularly problematic because, as we will see later, very small values of \( \gamma \) are often avoided in practice due to numerical stability issues.

In view of this discussion, our objective is to replace the MY envelope with a new envelope that has the same minimizers as \( F \) for all sufficiently small \( \gamma \), not just in the limit of \( \gamma \to 0 \).

### 4 Forward–backward envelope

In this section, we will study an envelope that addresses the limitations of the MY envelope. More specifically, for \( \gamma > 0 \), let us recall from Stella et al. (2017) that the forward–backward (FB) envelope of the function \( F \) in (6b) is defined
as

\[ F_{\gamma}^{FB}(x) := f(x) - \frac{\gamma}{2} \| \nabla f(x) \|^2 + g_{\gamma}(x - \gamma \nabla f(x)), \quad \text{(FB)} \]

where \( g_{\gamma} \) was defined in (8). A number of useful properties of \( F_{\gamma}^{FB} \) are collected below for the convenience of the reader, borrowed from Stella et al. (2017). Recall that \( P_{\gamma g} \) in (9) denotes the proximal operator associated with the function \( \gamma g \).

**Proposition 2 (Properties of the FB Envelope)** Suppose that Assumption 1 is fulfilled. For \( \gamma \in (0, 1/\lambda_2) \) and every \( x \in \mathbb{R}^d \), it holds that

1. \( F(P_{\gamma g}(x - \gamma \nabla f(x))) \leq F_{\gamma}^{FB}(x) \leq F(x) \), which relates the function \( F \) to its FB envelope.
2. \( F_{\gamma}^{MY}(x) \leq F_{\gamma}^{FB}(x) \leq F_{\gamma}^{MY}(x) \), which relates the MY and FB envelopes of the function \( F \).
3. \( F_{\gamma}^{FB} \) is continuously differentiable and its gradient is given by

\[ \nabla F_{\gamma}^{FB}(x) = \left(1 - \gamma \nabla^2 f(x)\right) \left(x - P_{\gamma g}(x - \gamma \nabla f(x))\right). \]

4. \( \arg \min F_{\gamma}^{FB} = \arg \min F \), i.e., the function \( F \) and its FB envelope have the same minimizers.

In view of Proposition 2.4, a remarkable property of the FB envelope is that the modes of the distribution \( p_{\gamma}^{FB} \propto \exp(-F_{\gamma}^{FB}) \) coincide with the modes of the target distribution \( p \propto \exp(-F) \), for all sufficiently small \( \gamma \), rather than only in the limit of \( \gamma \to 0 \). Indeed, very small values of \( \gamma \) are often avoided in practice due to numerical stability issues. This observation signifies the advantage of the FB envelope over the MY envelope. Recall that the modes of \( p_{\gamma}^{MY} \) coincide with those of \( p \) only in the limit of \( \gamma \to 0 \); see Sect. 3.

As a side note, let us remark that the proximal gradient descent algorithm for minimizing the (non-smooth) function \( F \) coincides with the gradient descent (with variable metric) for minimizing the (smooth) function \( F_{\gamma}^{FB} \), whenever \( \gamma \) is sufficiently small (Stella et al. 2017). It is also easy to use Proposition 2 to check the admissibility of the FB envelopes, as summarized below.

**Proposition 3 (Admissibility of FB Envelopes)** Suppose that Assumption 1 is fulfilled. Then \( \{F_{\gamma}^{FB} : \gamma \in (0, \gamma^{FB})\} \) are admissible envelopes of \( F \) in (6b), where

\[ \gamma^{FB} := \frac{1}{2\lambda_2 + 2\lambda_3(\lambda_0 + R)}. \quad \text{(10)} \]

Moreover,

\[ \left\| \nabla F_{\gamma}^{FB}(x) - \nabla F_{\gamma}^{FB}(y) \right\|_2 \leq \lambda_{\gamma}^{FB} \| x - y \|_2, \quad \text{(11a)} \]

\[ \langle x - y, \nabla F_{\gamma}^{FB}(x) - \nabla F_{\gamma}^{FB}(y) \rangle \geq \mu_{\gamma}^{FB} \| x - y \|_2^2, \quad \text{(11b)} \]

for every \( x, y \in \mathbb{R}^d \). The second inequality above holds when \( \| x - y \|_2 \geq \rho_{\gamma}^{FB} \). Lastly,

\[ \lambda_{\gamma}^{FB} := \gamma^{-1} + 2\lambda_2 + 3(\lambda_0 + R), \]

\[ \mu_{\gamma}^{FB} := \lambda_2 + 3(\lambda_0 + R), \]

\[ \rho_{\gamma}^{FB} := \frac{2R}{1 - 2\gamma(\lambda_2 + 3(\lambda_0 + R))}. \]

The Eq. (11) provides valuable information about the landscape of the FB envelope of \( F \), which we now summarize: (11a) means that \( F_{\gamma}^{FB} \) is a \( \lambda_{\gamma}^{FB} \)-smooth function. The smoothness of \( F_{\gamma}^{FB} \) in (FB) is not surprising since both \( f \) and \( g_{\gamma} \) are smooth functions. (Recall that \( g_{\gamma} \) is the MY envelope of \( g \), which is known to be \( \gamma^{-1} \)-smooth.) Moreover, even though \( F_{\gamma}^{FB} \) is not necessarily a strongly convex function, (11b) implies that \( F_{\gamma}^{FB} \) behaves like a strongly convex function over long distances. As detailed in the proof, (11b) holds essentially because the MY envelope of the indicator function \( 1_K \) is the “quadratic” function \( \frac{1}{2\gamma} \| x \|_2^2 \). The latter function grows quadratically far away from the origin. Here, \( \text{dist}(\cdot, K) \) is the distance to the set \( K \).

It is worth noting that a similar result to Proposition 3 is implicit in Brosse et al. (2017) for the MY envelope. That is, the MY envelope \( F_{\gamma}^{MY} \) also satisfies (11), albeit with different constants.

**Remark 3 (Convergence in the Wasserstein Metric)** Recall from Remark 1 that \( p_{\gamma}^{FB} \) converges weakly to \( p \) in the limit of \( \gamma \to 0 \). This weak convergence implies convergence in the Wasserstein metric by Bou-Rabee and Eberle (2020, Lemma 2.6), i.e.,

\[ \lim_{\gamma \to 0} W_1(p_{\gamma}^{FB}, p) = 0. \quad \text{(12)} \]

We recall here that, for two probability measures \( q_1 \) and \( q_2 \) that satisfy \( E_{q_1} \| x \|_2 < \infty \) and \( E_{q_2} \| y \|_2 < \infty \), their 1-Wasserstein or Kantorovich distance (Villani 2009) is

\[ W_1(q_1, q_2) := \inf_{x \sim q_1, y \sim q_2} E_{\text{dist}} x - y \|_2. \quad \text{(13)} \]

With some abuse of notation, throughout this work, we will occasionally replace the probability measures above with the corresponding probability distributions or random variables.
A non-asymptotic version of Remark 3 is presented below, which bounds the Wasserstein distance between the two probability measures \( p^{\text{FB}}_\gamma \) and \( p \). The result below is an analogue of Brosse (2017, Proposition 5) for the MY envelope. The key ingredient of their result is the Steiner’s formula for the volume of the set \( \set K + B(0, t) = \{ x : \text{dist}(x, K) \leq t \} \), where the sum is in the Minkowski sense. Essentially, our proof strategy is to use Proposition 2.2 to relate the FB and MY envelopes and then invoke Brosse (2017, Proposition 5). Below needs update.

**Theorem 1 (Wasserstein distance between \( p^{\text{FB}}_\gamma \) and \( p \): Suppose that Assumption 1 is fulfilled. For \( \gamma \in (0, \gamma^{\text{FB}}) \), it holds that**

\[
W_1(p^{\text{FB}}_\gamma ; p) \leq c_1 + c_2 R_1(\gamma) + c_3 I_2(\gamma) + c_4 R, \tag{14}
\]

where

\[
c_1 := \frac{e^{\max_{x \in K} g(x) - \gamma \lambda / 2} - \min f(x) \int_{\|x\| = 2} e^{-f(x)} \frac{\text{dist}(x, K)^2}{2} \, dx}{\int e^{-f(x)} \frac{\text{dist}(x, K)^2}{2} \, dx} - \frac{e^{\max f(x)} - \min f(x)}{\int e^{-f(x)} \frac{\text{dist}(x, K)^2}{2} \, dx},
\]

\[
c_2 := e^{\max f(x) - \min f(x)},
\]

\[
c_3 := e^{\max f(x) - \min f(x) + 2 \max_{x \in K} g(x)},
\]

\[
c_4 := e^{\max_{x \in K} g(x) - \min_{x \in K} g(x)} - e^{\min_{x \in K} g(x) - \max_{x \in K} g(x)}, \tag{15}
\]

\[
I_1(\gamma) := \sum_{i=0}^{d-1} d_i(K) \cdot (2\pi \gamma)^{\frac{d-i}{2}},
\]

\[
I_2(\gamma) := \sum_{i=0}^{d-1} d_i(K) \cdot (2\pi \gamma)^{\frac{d-i}{2}} \left( \sqrt{\gamma(d-i + 3)} + R \right).
\]

Above, \( d_i(K) \) is the \( i \)-th intrinsic volume of \( K \); see Klain and Rota (1997). In particular, the \( d \)-th volume of \( K \) coincides with the standard volume of \( K \), i.e., \( d_d(K) = d(K) \). Moreover, to keep the notation light, above we have suppressed the dependence of \( c_1 \) to \( c_5 \) on \( K, f, g, \gamma \).

As a sanity check, consider the special case of \( g = 1_K \), where \( 1_K \) is the indicator function for the set \( K \). Then we can use (15) to verify that \( c_1 \) and \( I_1(\gamma) \) and \( I_2(\gamma) \) and \( I_2(\gamma/(1 - \gamma \lambda / 2)) \) vanish when we send \( \gamma \to 0 \). Consequently, both the left- and right-hand sides of (14) vanish if we send \( \gamma \to 0 \).

When \( g = 1_K \), then Theorem 1 is precisely the analogue of Brosse (2017, Proposition 5) for the FB envelope. Their work, however, does not cover the case of \( g \neq 1_K \).

When \( g \neq 1_K \) and \( \gamma \to 0 \), the right-hand side of (14) converges to the nonzero value

\[
\left( e^{2\max_{x \in K} g(x)} - e^{-2\max_{x \in K} g(x)} \right) \int_{\|x\|=2} e^{-f(x)} \, dx
+
\left( e^{\min_{x \in K} g(x)} - e^{\max_{x \in K} g(x)} \right),
\]

unlike the left-hand side of (14), which converges to zero by (12). Improving (14) in the case \( g \neq 1_K \) appears to require highly restrictive assumptions on \( g \) which we wish to avoid here. Loosely speaking, the technical difficulty can be described as follows: As the volume is obtained by integrating the indicator function, the Steiner formula is particularly suited for the case \( g = 1_K \). When \( g \neq 1_K \), we are not aware of an analogue of the Steiner’s formula. At any rate, very small values of \( \gamma \) are often avoided in practice due to numerical stability issues. In this sense, improving (14) for very small values of \( \gamma \) might have limited practical value.

To summarize this section, we established that the FB envelopes \( \{ F^{\text{FB}}_\gamma : \gamma \in (0, \gamma^{0,\text{FB}}) \} \) are admissible and we can thus use them as a differentiable proxy for the non-smooth function \( F \) in (6b). Crucially, the FB envelope addresses the key limitation of the MY envelope, i.e., the modes of \( p^{\text{FB}}_\gamma \) coincide with the modes \( p \propto \exp(-F) \), for all sufficiently small \( \gamma \), rather than only in the limit of \( \gamma \to 0 \).

5 EULA: envelope unadjusted Langevin algorithm

We have so far introduced two smooth envelopes for the non-smooth function \( F \) in (6b), namely, the MY envelope \( F^{\text{MY}}_\gamma \) in (MY) and the FB envelope \( F^{\text{FB}}_\gamma \) in (FB). We also described in Sect. 4 the advantage of the FB envelope over the MY envelope. To keep our discussion general, below we consider admissible envelopes \( \{ F_\gamma : \gamma \in (0, \gamma^{0}) \} \) for the target function \( F \) in (6b); see Definition 1. Our discussion below can be specialized to either of the envelopes by setting \( F_\gamma = F^{\text{MY}}_\gamma \) or\( F_\gamma = F^{\text{FB}}_\gamma \), respectively.

For the time being, let us fix \( \gamma \in (0, \gamma^{0}) \). Unlike \( F \), note that \( \nabla F_\gamma \) exists and is Lipschitz continuous by Definition 1.3. We can now use the ULA (Dalalyan and Karagulyan 2019) to sample from \( p_\gamma \propto \exp(-F_\gamma) \), as a proxy for the target measure \( p \propto \exp(-F) \). The \( k \)-th iteration of the resulting algorithm is

\[
x_{k+1} = x_k - h
\nabla F_\gamma(x_k) + \sqrt{2h} \xi_{k+1}, \tag{16}
\]
where $h$ is the step size and $\zeta_{k+1} \in \mathbb{R}^d$ is a standard Gaussian random vector, independent of $\{\zeta_i\}_{i \leq k}$. In particular, if we choose $F_y = F_{\text{MY}}$, then (16) coincides with the MYULA from Brosse et al. (2017).

Under standard assumptions, to be reviewed later, the Markov chain $\{x_k\}_{k \geq 0}$ in (16) has a unique invariant probability measure, which we denote by $\hat{\pi}_{y,h}$. There are two sources of error that contribute to the difference between $\hat{\pi}_{y,h}$ and the target measure $\pi$ in (6a), which we list below:

1. First, note that (16) is only intended to sample from $p_y$, as a proxy for the target distribution $p$. That is, the first source of error is the difference between the two probability measures $p_y$ and $p$.

2. Second, the step size $h$ is known to contribute to the difference between the two probability measures $\hat{\pi}_{y,h}$ and $p_y$; see Dalalyan and Karagulyan (2019). This bias vanishes only in the limit of $h \to 0$.

In fact, instead of (16), we study here a slightly more general algorithm that allows $\gamma$ and $h$ to vary. More specifically, for a nonincreasing sequence $\{\gamma_k\}_{k \geq 0}$ and step sizes $\{h_k\}_{k \geq 0}$, the $k$-th iteration of this more general algorithm is

$$x_{k+1} = x_k - h_k \nabla F_{\gamma_k}(x_k) + \sqrt{2h_k} \zeta_{k+1}, \quad (\text{EULA})$$

where $\zeta_{k+1} \in \mathbb{R}^d$ is a standard Gaussian random vector independent of $\{\zeta_i\}_{i \leq k}$ (EULA) stands for Envelope Unadjusted Langevin Algorithm.

In particular, if we set $\gamma_k = \gamma$ in (EULA) for every $k \geq 0$, then we retrieve (16). Alternatively, if $\gamma_k \geq 0$ is a decreasing sequence, then $F_{\gamma_k}$ becomes an increasingly better approximation of the target potential function $F$ as $k$ increases; see Definition 1.2. That is, (EULA) uses increasingly better approximations of the potential function $F$ as $k$ increases.

We next present the iteration complexity of (EULA) for admissible envelopes $\{F_y : \gamma \in (0, \gamma^0)\}$, where admissibility was defined in Definition 1. The result below can be specialized to both MY and FB envelopes by setting $F_y = F_{\gamma}^{\text{MY}}$ or $F_y = F_{\gamma}^{\text{FB}}$, respectively.

**Theorem 2 (Iteration Complexity of (EULA))** For $\gamma > 0$, consider admissible envelopes $\{F_y : \gamma \in (0, \gamma^0)\}$ of $F$ in (6b); see Definition 1. For $\gamma > 0$ and $\rho_\gamma \geq 0$, we additionally assume that $F_y$ satisfies

$$\langle x - y, \nabla F_y(x) - \nabla F_y(y) \rangle \geq \mu_\gamma \|x - y\|^2_2, \quad (17)$$

when $\|x - y\|_2 \geq \rho_\gamma$ and $\gamma \in (0, \gamma^0)$. Consider two sequences $\{\gamma_k\}_{k \geq 0} \subset (0, \gamma^0)$ and $\{h_k\}_{k \geq 0} \subset \mathbb{R}_+$. For the algorithm (EULA), let $q_k$ denote the law of $x_k$ for every integer $k \geq 0$. That is, $x_k \sim q_k$ for every $k \geq 0$. Then the $W_1$ distance between $q_k$ and the target measure $\pi \propto e^{-F}$ in (6a) is bounded by

$$W_1(q_k, \pi) \leq e^6 e^{\rho_\gamma^2} \sum_{i=0}^{k-1} \prod_{j=i+1}^{k-1} (1 - c_h h_j) + W_1(q_0, \pi_0)$$

$$+ e^{h_\gamma^2} \sum_{i=0}^{k-1} \prod_{j=i+1}^{k-1} (1 - c_h h_j) + W_1(\pi_0, \pi),$$

for every $k \geq 0$, provided that

$$h_k \leq \min \left\{ \frac{1}{\lambda^{\gamma_k}}, \frac{\lambda^{\gamma_k} \rho^{2}_{\gamma_k}}{6}, \frac{c_0^2}{970 \lambda^{\gamma_k} \rho^{2}_{\gamma_k}} \right\}.$$  \hspace{1cm} (18)

Above, $c_0 \geq 0.007$ is a universal constant specified in Eberle and Majka (2019, Eq. (6.6)). Moreover,

$$c_6 := (1 + h_k \lambda^{\gamma_k}) \rho^{2}_{\gamma_k} \leq 7 \rho^{2}_{\gamma_k}/6, \quad c_7 := 7 \lambda^{\gamma_k} \rho^{2}_{\gamma_k}/c_0,$$

$$c_8 := \min \left\{ \frac{\lambda^{\gamma_k} \rho^{2}_{\gamma_k}}{24 \gamma^0}, \frac{245}{24 \gamma^0} \right\} (\lambda^{\gamma_k} \rho^{2}_{\gamma_k})^2 e^{- \frac{6 \gamma^0}{\gamma^0} \lambda^{\gamma_k} \rho^{2}_{\gamma_k}},$$

$$\omega_k := \sqrt{\lambda^{\gamma_k} h_k^2 d \cdot \left( \sqrt{h_k \lambda^{\gamma_k}} + \rho_\gamma \right)} + W_1(\pi, \pi_0) \leq W_1(\pi_0, \pi), \quad k \geq 0.$$

When $\rho_\gamma = 0$, then note that (17) requires $F_y$ to be $\mu_\gamma$-strongly convex for every $\gamma \in (0, \gamma^0)$. This can happen, for example, when $f$ itself is a strongly convex function. A shortcoming of Theorem 2 is that $h_k = 0$ if $\rho_\gamma = 0$. It becomes clear from our proof that we have inherited this shortcoming from Eberle and Majka (2019); see Proposition 4. Their work also offers a more involved result, which does not suffer from this issue; see Eberle and Majka (2019, Theorem 2.12). However, we opted in this work for their simplified result in Proposition 4, though this can be improved.

A particularly important special case of Theorem 2 is when we choose the FB envelope, and use a fixed $\gamma$ and step size $h$. 

**Corollary 1 (Iteration Complexity of (EULA) for FB Envelope)** Suppose that Assumption 1 is fulfilled. For the algorithm (EULA), suppose that $\gamma_k = \gamma \in (0, \gamma^0)$ and $F_y = F^{\gamma}$ and $h_k = h > 0$ for every integer $k \geq 0$, see (FB) and (10). In (EULA), also let $q_k$ denote the law of $x_k$ for every $k \geq 0$. That is, $x_k \sim q_k$ for every integer $k \geq 0$. Then the $W_1$ distance between $q_k$ and the target measure $\pi \propto e^{-F}$ in (6a) is bounded by

$$W_1(q_k, \pi) \leq e^6 e^{\rho_\gamma^2} (1 - c_h h)^k W_1(q_0, \pi_0)$$

$$+ \frac{\alpha_{FB} e^{h_\gamma^2} (1 - c_h h)^k + c_5}{c_8 h},$$  \hspace{1cm} (19)
for every $k \geq 0$, provided that $\gamma \in (0, \gamma^0)$, and
\[
 h_k \leq \frac{1}{\lambda \gamma} \min \left( \frac{1}{6} \frac{\mu^F_B}{\lambda \gamma}, \frac{\lambda \gamma (\rho^F_B)^2}{3 \cdot 970 \lambda \gamma (\rho^F_B)^2} \right).
\]

Above, $c_0 \geq 0.007$ is a universal constant specified in Eberle and Majka (2019, Eq. (6.6)). Moreover,
\[
c^F_B := (1 + h_k \lambda \gamma)(\rho^F_B)^2 \leq \frac{7}{6} \rho^F_B / 6,
\]
\[
c^F_B := 7 h_k \lambda \gamma / c_0,
\]
\[
c^F_B := \min \left( \frac{\mu^F_B}{2}, \frac{245}{24 c_0} \frac{(\lambda \gamma \rho^F_B)^2}{(\mu^F_B)^2} \right) e^{-\frac{\lambda \gamma}{2 c_0} (\rho^F_B)^2},
\]
\[
a^F_B := \lambda \gamma \sqrt{h^2} \cdot \left( \sqrt{h^2} \lambda \gamma + \sqrt{2} \right) + 2 c^F_B.
\]
The remaining quantities were defined in Propositions 3 and 1.

We remark that Corollary 1 for the FB envelope is the analogue of Brosse (2017, Proposition 7) for the MY envelope. However, note that Brosse (2017, Proposition 7) requires $f$ to be strongly convex whereas we merely assume $f$ to be convex, see Assumption 1.

## 6 Proof of Theorem 2

To begin, we let $Q_k$ denote the Markov transition kernel associated with the Markov chain $\{x_k\}_{k \geq 0}$. This transition kernel is specified as
\[
Q_k(x, \cdot) := \text{Normal}(x - h_k \nabla F_{\gamma}(x), 2 h_k I_d),
\]
(20)
where $\text{Normal}(\mu, B)$ is the Gaussian probability measure with mean $\mu \in \mathbb{R}^d$ and covariance matrix $B \in \mathbb{R}^{d \times d}$. Above, note that $Q_k$ depends on both $h_k$ and $\gamma_k$. We also let $q_k$ denote the law of $x_k$, i.e., $x_k \sim q_k$. Using the standard notation, we can now write that
\[
q_{k+1} = q_k Q_k.
\]
To be precise, (21) is equivalent to
\[
q_{k+1}(d y) = \int q_k(d x) Q_k(x, d y),
\]
(22)
for every $y \in \mathbb{R}^d$. Recall that $p_{k+1}$ serves as a proxy for the target probability measure $p$. The $W_1$ distance between $q_{k+1}$ and $p_{k+1}$ can be bounded as
\[
W_1(q_{k+1}, p_{k+1}) = W_1(q_k Q_k, p_{k+1}) \quad \text{(see (21))}
\]
\[
\leq W_1(q_k Q_k, p_{k+1}) + W_1(p_{k+1} Q_k, p_{k+1}),
\]
(23)
where $(i)$ uses the triangle inequality. We separately control each $W_1$ distance in the last line above. For the first distance, we plan to invoke Theorem 2.12 from Eberle and Majka (2019), reviewed below for the convenience of the reader. It is worth noting that a similar result to the one below appears in Majka et al. (2020, Corollary 2.4).

**Proposition 4** (Eberle and Majka 2019, THEOREM 2.12) Let
\[
c_6 := (1 + h_k \lambda \gamma)(\rho^F_B)^2 \leq \frac{7}{6} \rho^F_B / 6,
\]
\[
c_7 := 7 h_k \lambda \gamma / c_0,
\]
\[
\theta(r) := \int_0^r e^{-c_7 \min(s, r)} \, ds, \quad \Theta(x, y) := \theta(\|x - y\|_2),
\]
\[
c_8 := \min \left( \frac{\mu^F_B}{2}, \frac{245}{24 c_0} \frac{(\lambda \gamma \rho^F_B)^2}{(\mu^F_B)^2} \right) e^{-\frac{\lambda \gamma}{2 c_0} (\rho^F_B)^2},
\]
\[
h_k \leq \frac{1}{\lambda \gamma} \min \left( \frac{1}{6} \frac{\mu^F_B}{\lambda \gamma}, \frac{\lambda \gamma \rho^F_B}{3 \cdot 970 \lambda \gamma \rho^F_B} \right),
\]
where $c_0 \geq 0.007$ is a universal constant specified in Eberle and Majka (2019, Eq. (6.6)). Then it holds that
\[
W_\Theta(q_k Q_k, p_{k+1} Q_k) \leq (1 - c_8 h_k) \cdot W_\Theta(q_k, p_{k+1}).
\]
(24)
where $W_\Theta$ is defined similar to (13) but the $\ell_2$-norm is replaced with $\Theta$. Above, to keep the notation light, we have suppressed the dependence of $c_6$ and $c_7$ and $c_8$ on $K, f, g, h_k$. Moreover, the two metrics $W_\Theta$ and $W_1$ are related as
\[
e^{-c_6 c_7} W_1 \leq W_\Theta \leq W_1.
\]
(25)
As for the second $W_1$ distance in the last line of (23), the following result is standard; see the appendix for its proof.

**Lemma 1** (DISCRETIZATION ERROR) It holds that
\[
W_1(p_{k+1} Q_k, p_{k+1}) \leq c_9 := \lambda \gamma \sqrt{h^2} \cdot \left( \sqrt{h^2} \lambda \gamma + \sqrt{2} \right).
\]
(26)
In fact, it is more common to write the left-hand side of (26) in terms of the Markov transition kernel of the corresponding Langevin diffusion, as discussed in the proof of Lemma 1. By combining Proposition 4 and Lemma 1, we can now revisit (23) and write that
\[
W_\Theta(q_{k+1}, p_{k+1}) \leq W_\Theta(q_k Q_k, p_{k+1} Q_k) + c_9,
\]
\[
\leq (1 - c_8 h_k) W_\Theta(q_k, p_{k+1}) + W_1(p_{k+1} Q_k, p_{k+1}),
\]
\[
\leq (1 - c_8 h_k) W_\Theta(q_k, p_{k+1}) + c_9,
\]
(27)
where $(i)$ uses (23), $(ii)$ uses Proposition 4 and (25), and $(iii)$ uses Lemma 1. Using the triangle inequality, it immediately
follows that

\[ W_\Theta(q_{k+1}, \rho_{k+1}) \]
\[ \leq W_\Theta(q_{k+1}, \rho_{k+1}) + W_\Theta(\rho_{k+1}, \rho) + W_\Theta(\rho_{k+1}, \rho) \]
\[ \leq (1 - c q_h) W_\Theta(q_k, \rho_{k+1}) + c q + W_\Theta(q_k, \rho) + W_\Theta(\rho_{k+1}, \rho) \]
\[ =: (1 - c q_h) W_\Theta(q_k, \rho_{k+1}) + \alpha_k. \quad (28) \]

where (i) uses the triangle inequality, (ii) uses (27), and (iii) uses (25). By unwrapping (28), we find that

\[ W_1(q_k, \rho_{k+1}) \]
\[ \leq e^{c q_r} W_\Theta(q_k, \rho_{k+1}) \]
\[ \leq e^{c q_r} \prod_{i=0}^{k-1} (1 - c q h_i) \cdot W_\Theta(q_0, \rho_{k+1}) + \alpha_k \]
\[ \leq e^{c q_r} \prod_{i=0}^{k-1} (1 - c q h_i) \cdot W_1(q_0, \rho_{k+1}) + e^{c q_r}, \quad (29) \]

where (i) and (iii) use (25), and (ii) uses (27). Lastly, we can use (29) in order to bound the \( W_1 \) distance at iteration \( k \) to the target measure \( \rho \) as

\[ W_1(q_k, \rho) \]
\[ \leq W_1(q_k, \rho_{k+1}) + W_1(\rho_{k+1}, \rho) \]
\[ \leq e^{c q_r} \prod_{i=0}^{k-1} (1 - c q h_i) \cdot W_1(q_0, \rho_{k+1}) + e^{c q_r}. \]
\[ \sum_{i=0}^{k-1} \alpha_i \prod_{j=i+1}^{k-1} (1 - c q h_j) + W_1(\rho_{k+1}, \rho), \quad (30) \]

where (i) uses the triangle inequality, and (ii) uses (27). This completes the proof of Theorem 2.

7 Numerical experiments

A few numerical experiments are presented below to support our theoretical contributions.

7.1 Truncated Gaussian

Our first numerical experiment deals with sampling from a truncated Gaussian distribution, restricted to a box \( K_d \subset \mathbb{R}^d \).

For this problem the potential \( U : \mathbb{R}^d \rightarrow \mathbb{R} \) is specified as

\[ U(x) := \frac{1}{2} \left( x \cdot \Sigma^{-1} x \right) + i K_d(x). \quad (31) \]

Here similarly to Brosse et al. (2017) the \((i, j)\)th entry of the covariance matrix is given by

\[ \Sigma_{i,j} := \frac{1}{1 + |i - j|}. \]

We now consider three scenarios, namely,

- \( d = 2 \) with \( K_2 = [0, 5] \times [0, 1] \),
- \( d = 10 \) with \( K_{10} = [0, 5] \times [0, 0.5]^9 \)
- \( d = 100 \) with \( K_{100} = [0, 5] \times [0, 0.5]^{99} \).

Using quadrature techniques, it is possible in the two-dimensional case \( (d = 2) \) to calculate exactly the mean and the covariance of the truncated Gaussian distribution, as well as the corresponding approximations obtained via MY and FB envelopes. More specifically, Fig. 1 uses MATLAB’s \texttt{integral2} command to plot the following quantities for various values of \( \gamma \):

\[ \mathbb{E}_{p^{\text{MY}}} [x_1] := \frac{\int_{-\infty}^{\infty} x_1 e^{-F_{p^{\text{MY}}}(x_1)} dx}{\int_{-\infty}^{\infty} e^{-F_{p^{\text{MY}}}(x_1)} dx} - (\mathbb{E}_{p^{\text{MY}}} [x_1])^2, \]
\[ \text{var}_{p^{\text{MY}}} [x_1] := \frac{\int_{-\infty}^{\infty} x_1^2 e^{-F_{p^{\text{MY}}}(x_1)} dx}{\int_{-\infty}^{\infty} e^{-F_{p^{\text{MY}}}(x_1)} dx} - (\mathbb{E}_{p^{\text{MY}}} [x_1])^2, \]
\[ \mathbb{E}_{p^{\text{FB}}} [x_1] := \frac{\int_{-\infty}^{\infty} x_1 e^{-F_{p^{\text{FB}}}(x_1)} dx}{\int_{-\infty}^{\infty} e^{-F_{p^{\text{FB}}}(x_1)} dx} - (\mathbb{E}_{p^{\text{FB}}} [x_1])^2, \]
\[ \text{var}_{p^{\text{FB}}} [x_1] := \frac{\int_{-\infty}^{\infty} x_1^2 e^{-F_{p^{\text{FB}}}(x_1)} dx}{\int_{-\infty}^{\infty} e^{-F_{p^{\text{FB}}}(x_1)} dx} - (\mathbb{E}_{p^{\text{FB}}} [x_1])^2. \]

The horizontal lines in Fig. 1 show the ground truth values obtained by MATLAB’s \texttt{integral2} command, i.e.,

\[ \mathbb{E}_{p} [x_1] := \frac{\int_{-\infty}^{\infty} x_1 e^{-F(x_1)} dx}{\int_{-\infty}^{\infty} e^{-F(x_1)} dx} - (\mathbb{E}_{p} [x_1])^2. \]
\[ \text{var}_{p} [x_1] := \frac{\int_{-\infty}^{\infty} x_1^2 e^{-F(x_1)} dx}{\int_{-\infty}^{\infty} e^{-F(x_1)} dx} - (\mathbb{E}_{p} [x_1])^2. \]

For small values of the parameter \( \gamma \), we observe in Fig. 1 that the FB envelope better approximates the mean of the first component than the MY envelope. However, the FB envelope tends to overestimate the variance. This can be understood by comparing the two envelopes since in the case of the MY
(a) $\mathbb{E}(x_1)$

(b) $\text{Var}(x_1)$

Fig. 1 This figure compares the MY and FB envelopes for the two-dimensional truncated Gaussian distribution $p \propto e^{-U - 1/k_2}$ specified in Sect. 7.1. The horizontal lines in the top and bottom panels show, respectively, the expectation and variance of the first coordinate, namely, $\mathbb{E}_p[x_1]$ and $\text{Var}_p[x_1] = \mathbb{E}_p[x_1^2] - (\mathbb{E}_p[x_1])^2$. The blue and red graphs in both panels show the estimated values of $\mathbb{E}_p[x_1]$ and $\text{Var}_p[x_1]$, obtained via MY and FB envelopes. That is, the top graphs correspond to $\mathbb{E}_{p^{\gamma}}[x_1]$ and $\text{Var}_{p^{\gamma}}[x_1]$, for various values of $\gamma$. Similarly, the bottom graphs correspond to $\text{Var}_{p^{\gamma}}[x_1]$ and $\text{Var}_{p^{\gamma}}[x_1]$, for various values of the parameter $\gamma$.

Such explicit calculations are not tractable in higher dimensions, i.e., for $d \in \{10, 100\}$. Instead, we now generate $10^6$ samples from the truncated Gaussian distribution $p$ by applying MYULA and FBULA. As our ground truth, we also generate $10^5$ samples from $p$ with the wall HMC (wHMC) (Pakman and Paninski 2014). In all three approaches, the initial 10% of the obtained samples are discarded as the burn-in period, while each experiment is repeated 100 times. In terms of the parameters, we set $\gamma = 0.05$ and fix $h = 0.005$ for all of our experiments.

The results are visualized in Figs. 2 and 3. More specifically, Fig. 2 corresponds to $d = 10$ and shows the estimates for $\mathbb{E}_p[x_i]$ for $i \in \{1, 2, 3\}$, obtained by MYULA, FBULA, and wHMC. Similarly, Fig. 3 corresponds to $d = 100$. These figures indicate that, in all of these cases, FBULA is providing a more accurate approximation of the expectation compared to MYULA.
7.2 Tomographic image reconstruction

We now study a tomographic image reconstruction problem. In this case the true image is the Shepp–Logan phantom test image of dimension $d = 128 \times 128$ (Shepp and Logan 1974), in which we applied a Fourier operator $F$ followed by a subsampling operator $A$, reducing the observed pixels by approximately 85%. Finally, zero-mean additive Gaussian noise $\xi$ is added with standard deviation $\sigma = 10^{-2}$ to produce an incomplete observation $y = AFx + \xi$ where $y \in \mathbb{C}^p$. Note that $p < d$. With regards to the prior, we use the total-variation norm with an additional constraint for the size of the pixels. This leads to the following posterior distribution:

$$
\pi(x) \propto \exp \left[ -\frac{\|y - AFx\|^2}{2\sigma^2} - \beta \text{TV}(x) - 1_{[0,1]^d}(x) \right],
$$

with $\beta = 100$. Above, $1_{[0,1]^d}$ is the convex indicator function on the unit cube, as the pixel values for this experiment are scaled to the range $[0, 1]$. Here $\text{TV}(x)$ represents the

---

**Fig. 4** Tomography experiment: a True image $x$ of dimension $d = 128 \times 128$. b Incomplete and noisy observation $y$, amplitude of Fourier coefficients in logarithmic scale.

**Fig. 5** Tomography experiment: a Convergence to the typical set of the posterior distribution (32). b Evolution of the MSE in stationarity.
total-variation pseudo-norm (Rudin et al. 1992; Chambolle 2004). Following (4) and (6b), we have that \( f(x) = \|y - AFx\|^2/2\sigma^2 \) and \( g(x) = \beta TV(x) + 1_{[0,1]^d}(x) \).

Figure 4a shows the Shepp–Logan phantom tomography test image for this experiment and Fig. 4b shows the amplitude of the (noisy) Fourier coefficients collected in the observation vector \( y \) (in logarithmic scale). In this figure, black regions represent unobserved pixels.

We have set \( \gamma = 1/(5L_f) \) where \( L_f = 1/\sigma^2 = 10^4 \) for both MYULA and FBULA. Fig. 5a shows the evolution of the values of \( \log \pi(x) \) from (32) of both MYULA and FBULA with the step-size \( h = 1/(L_f + 1/\gamma) = 1.67 \times 10^{-5} \). We observe that both methods converge at a similar rate in terms of the actual computational time required to run the algorithms. However, Fig. 5b shows the evolution of the mean-squared error (MSE) between the ergodic mean of the samples and the true image \( x \). Here, it can be seen FBULA reaches a better MSE level compared to MYULA. We have also included in Fig. 6a and b the posterior mean estimated by both MYULA and FBULA.

Acknowledgements The authors would like to thank Marcelo Pereyra for valuable comments on a draft manuscript. The work of KCZ was supported by the Leverhulme Trust (RF/ 2020-310) and by the Engineering and Physical Sciences Research Council under grant EP/V006177/1. The work of AE was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

Appendix A: Proof of Proposition 1

To establish the admissibility of the MY envelope, we verify the conditions in Definition 1. We begin by verifying that Definition 1.3 is met. Recall that the MY envelope is continuously-differentiable Rockafellar and Wets (2009, Theorem 2.26), and its gradient is given by

\[
\nabla F^{\text{MY}}_{\gamma}(x) = \nabla f(x) + \frac{x - P_{\gamma g}(x)}{\gamma},
\]

\( P_{\gamma g}(x) = \arg \min_{z \in \mathbb{R}^d} \left\{ g(z) + \frac{1}{2\gamma} \|x - z\|^2 \right\}. \) (A2)

Here, \( P_{\gamma g} \) is the proximal operator associated with the function \( \gamma g \). Above, note that the minimizer is unique and the map \( P_{\gamma g} \) is thus well-defined. Recall also that we can use the Moreau decomposition (Beck 2017) to write that

\[
x = P_{\gamma g}(x) + P_{g^*(\cdot/\gamma)}(x), \quad x \in \mathbb{R}^d,
\]

where \( g^* \) is the Fenchel conjugate of \( g \). Using (A3), we rewrite (A1) as

\[
\nabla F^{\text{MY}}_{\gamma}(x) = \nabla f(x) + \gamma^{-1} P_{\gamma g^*(\cdot/\gamma)}(x),
\]
which we will use next to compute the Lipschitz constant of \( \nabla F_{\gamma}^{\text{MY}} \). For \( x, y \in \mathbb{R}^d \), note that
\[
\| \nabla F_{\gamma}^{\text{MY}}(x) - \nabla F_{\gamma}^{\text{MY}}(y) \|_2 \\
\leq \| \nabla f(x) - \nabla f(y) \|_2 \\
+ \gamma^{-1} \| P_{\gamma g^*(\cdot/\gamma)}(x) - P_{\gamma g^*(\cdot/\gamma)}(y) \|_2 \\
\leq (\lambda_2 + \gamma^{-1}) \| x - y \|_2. \tag{A5}
\]

where \((i)\) uses the triangle inequality and \((A4)\), and \((ii)\) uses \((3c)\) and the non-expansiveness of the proximal operator. We next verify that Definition 1.2 holds. In one direction, it is easy to see from \((8)\) that \( g_{\gamma} \leq g \) for every \( \gamma > 0 \). In the other direction, we fix \( x \in \mathbb{R}^d \) and distinguish two cases:

1. When \( x \in K \), let us fix an arbitrary \( \epsilon > 0 \). Because \( g(x) < \infty \) and \( \min_{z} g(z) = -\infty \) by \((5)\), there exists a sufficiently small \( \gamma_{\epsilon} > 0 \) such that the following holds for every \( \gamma \leq \gamma_{\epsilon} \):
\[
\min_{z \in \mathbb{R}^d} \left\{ g(z) + \frac{\epsilon^2}{2\gamma} \right\} > g(x). \tag{A6}
\]

We now use the above inequality as part of the following argument,
\[
\min_{z \in \mathbb{R}^d} \left\{ g(z) + \frac{1}{2\gamma} \| x - z \|_2^2 : \| x - z \| \geq \epsilon \right\} \\
\geq \min_{z} \left\{ g(z) + \frac{\epsilon^2}{2\gamma} \right\} \\
\geq g_{\gamma}(x) = \min_{z} \left\{ g(z) + \frac{1}{2\gamma} \| x - z \|_2^2 \right\}, \tag{A7}
\]

which holds when \( x \in K \) and for every \( \gamma \leq \gamma_{\epsilon} \). Above, \((i)\) uses \((A6)\). In view of \((A2)\) and \((A7)\), we conclude that \( \| x - P_{\gamma g}(x) \|_2 \leq \epsilon \), for every \( x \in K \) and every \( \gamma \leq \gamma_{\epsilon} \).

Since the choice of \( \epsilon \) was arbitrary, we arrive at
\[
\lim_{\gamma \to 0} \| x - P_{\gamma g}(x) \|_2 = 0, \quad x \in K. \tag{A8}
\]

It immediately follows that
\[
\lim_{\gamma \to 0} g_{\gamma}(x) - g(x) \\
\leq \lim_{\gamma \to 0} \min_{z} \left\{ g(z) + \frac{1}{2\gamma} \| x - z \|_2^2 \right\} - g(x) \\
\leq \lim_{\gamma \to 0} \left\{ g(P_{\gamma g}(x)) + \frac{1}{2\gamma} \| x - P_{\gamma g}(x) \|_2^2 \right\} - g(x) \\
\geq \lim_{\gamma \to 0} g(P_{\gamma g}(x)) - g(x) \\
\geq \lim_{\gamma \to 0} \inf \{ g(z) - g(x) \} \\
\geq 0, \quad x \in K, \quad \text{provided that the maximum above is finite by \((5)\).}
\]

where \((i)\) uses \((MY)\), \((ii)\) uses \((A8)\), and \((i) \) uses the lower semi-continuity of \( g \) is lower semi-continuous; see before Eq. \((5)\). If we combine \((A9)\) with the earlier observation that \( g_{\gamma} \leq g \) for every \( \gamma \), we reach the conclusion that \( \lim_{\gamma \to 0} g_{\gamma}(x) = g(x) \), provided that \( x \in K \). Lastly, after recalling the definitions of \( F \) and \( F_{\gamma}^{\text{MY}} \) in \((6b)\) and \((MY)\), respectively, we arrive at
\[
\lim_{\gamma \to 0} F_{\gamma}^{\text{MY}}(x) = F(x), \quad x \in K. \tag{A10}
\]

2. When \( x \notin K \), we write that
\[
g_{\gamma}(x) = \min_{z} \left\{ g(z) + \frac{1}{2\gamma} \| x - z \|_2^2 \right\} \\
\leq \min_{z} \left\{ g(z) + \frac{1}{2\gamma} \| x - z \|_2^2 : z \in K \right\} \\
\geq - \max_{z \in K} g(z) + \frac{1}{2\gamma} \min_{z \in K} \| x - z \|_2^2 \\
=: - \max_{z \in K} g(z) + \frac{1}{2\gamma} \text{dist}(x, K)^2. \tag{A11}
\]

where \( \text{dist}(x, K) \) is the Euclidean distance from \( x \) to the set \( K \). Above, \((i)\) uses the fact that \( g \) is infinite outside the set \( K \) by \((5)\). Also, \((ii)\) uses the inequality \( \min h_1 + h_2 \geq \min h_1 - \max h_2 \) for real-valued functions \( h_1 \) and \( h_2 \). The maximum above is finite by \((5)\). By sending \( \gamma \) to zero above, we immediately find that \( \lim_{\gamma \to 0} g_{\gamma}(x) = \infty = g(x) \), provided that \( x \notin K \). Recalling the definitions of \( F \) and \( F_{\gamma} \) in \((6b)\) and \((MY)\), respectively, we conclude that
\[
\lim_{\gamma \to 0} F_{\gamma}^{\text{MY}}(x) = F(x) = \infty, \quad x \notin K. \tag{A12}
\]

Together, \((A8)\) and \((A12)\) imply that \( \lim_{\gamma \to 0} F_{\gamma}(x) = F(x) \) for every \( x \in \mathbb{R}^d \).

Lastly, we now verify that Definition 1.1 is satisfied: By Assumption 1.1, \( K \) is enclosed inside a ball of radius \( R \) centered at the origin, which implies that
\[
\text{dist}(x, K) \geq (\| x \| - R)_+, \quad x \in \mathbb{R}^d, \tag{A13}
\]

where \((a)_+ := \max(a, 0)\). When \( \| x \|_2 \) is sufficiently large, we can simplify \((A13)\). In particular, \((A13)\) immediately implies that
\[
\text{dist}(x, K) \geq \| x \|_2/2, \quad \| x \|_2 \geq 2R. \tag{A14}
\]
If \( \|x\|_2 \geq 2R \), then we can use the convexity of \( f \) in Assumption 1.2 to write that
\[
F^\text{MY}_\gamma(x) = f(x) + g_\gamma(x)
\]
(iii) \( \geq f(0) + \langle x, \nabla f(0) \rangle + g_\gamma(x) \)
(iv) \( \geq f(0) + \langle x, \nabla f(0) \rangle - \max_{z \in \mathbb{K}} g(z) + \frac{1}{2\gamma} \operatorname{dist}(x, \mathbb{K})^2 \)
\[
\geq f(0) + \langle x, \nabla f(0) \rangle - \max_{z \in \mathbb{K}} g(z) + \frac{\|x\|^2}{8\gamma}.
\]
for \( \|x\|_2 \geq 2R \). Above, (i) uses (MY), (ii) uses the convexity of \( f \), (iii) uses (A11), and (iv) uses (A14). We now set
\[
F^0(x) := \begin{cases} 
F^\text{MY}_\gamma(x), & \|x\|_2 \leq 2R \\
f(0) + \langle x, \nabla f(0) \rangle - \max_{z \in \mathbb{K}} g(z) + \frac{\|x\|^2}{8\gamma} & \|x\|_2 \geq 2R.
\end{cases}
\]
By its construction, note that \( F^0 \) satisfies \( \int e^{-F^0(x)} \, dx < \infty \) and \( F^\text{MY}_\gamma \geq F^0 \). That is, \( F^0 \) satisfies the requirements in Definition 1.1. The above claim about the finite integral is true because \( F^0(x) \) is quadratic for large \( \|x\|_2 \) and \( \exp(-F^0) \) thus decays rapidly far away from the origin. Having verified the requirements in Definition 1, the proof of Proposition 1 is now complete.

**Appendix B: Proof of Proposition 3**

To establish the admissibility of the FB envelope, we verify the requirements in Definition 1. We first verify that Definition 1.3 is satisfied. To begin, for \( x \in \mathbb{R}^d \), let
\[
\mathcal{X} := x - \gamma \nabla f(x), \quad T_\gamma(x) := P_{\gamma g}(\mathcal{X}),
\]
for short. Above, recall that \( P_{\gamma g} \) is the proximal operator associated with the function \( \gamma g \); see (9). We can then apply the Moreau decomposition (Beck 2017) to write that
\[
\mathcal{X} = P_{\gamma g}(\mathcal{X}) + P_{\gamma g^*(\gamma/\gamma)}(\mathcal{X}),
\]
where \( g^* \) is the Fenchel conjugate of \( g \). For future reference, we record the following observations:
\[
\|T_\gamma(x)\|_2 \leq \|P_{\gamma g}(\mathcal{X})\|_2 \leq R, \quad \|x - T_\gamma(x)\|_2 \leq \|x\|_2 + \|T_\gamma(x)\|_2 \leq \|x\|_2 + R, \quad \|x - y - (T_\gamma(x) - T_\gamma(y))\|_2 \leq \|x - y\|_2 + R,
\]
\[
\|x - y - (P_{\gamma g}(\mathcal{X}) - P_{\gamma g}(\mathcal{Y}))\|_2 \leq \|x - y\|_2 + R,
\]
\[
\|y(\nabla f(x) - \nabla f(y)) + (P_{\gamma g^*(\gamma/\gamma)}(\mathcal{X}) - P_{\gamma g^*(\gamma/\gamma)}(\mathcal{Y}))\|_2 \leq \gamma \|\nabla f(x) - \nabla f(y)\|_2 + \|P_{\gamma g^*(\gamma/\gamma)}(\mathcal{X}) - P_{\gamma g^*(\gamma/\gamma)}(\mathcal{Y})\|_2 \leq \gamma \|\nabla f(x) - \nabla f(y)\|_2 + \gamma \|\nabla f(x) - \nabla f(y)\|_2 \leq (1 + 2\gamma \lambda_2)\|x - y\|, \quad (B19)
\]
\[
\|I_d - \gamma \nabla^2 f(x)\|_2 \leq 1. \quad (B20)
\]
Above, \( I_d \in \mathbb{R}^{d \times d} \) is the identity matrix. Also, (i) combines Assumption 1.1 and the second line of (A11). Moreover, (ii) uses (B17), (iii) uses (B15), (iv) uses (B15) and (B16), (v) uses (3c) and the non-expansiveness of the proximal operator, (vi) uses the triangle inequality and (B15), and (vii) uses (3c). Lastly, (viii) uses (3c), and (ix) uses (B20) and the assumption that \( \gamma \in (0, 1/\lambda_2) \).

Recall the expression for \( \nabla F^\text{FB}(\gamma) \) from Proposition 2.3. We next establish that \( F^\text{FB}_\gamma \) is smooth. In our argument below, without loss of generality, we assume that \( \|y\|_2 \leq \lambda_0 \):
\[
\|\nabla F^\text{FB}_\gamma(x) - \nabla F^\text{FB}_\gamma(y)\|_2 = \|I_d - \gamma \nabla^2 f(x)(x - T_\gamma(x)) - (I_d - \gamma \nabla^2 f(y))(y - T_\gamma(y))\|_2 \leq \|I_d - \gamma \nabla^2 f(x)(x - y - (T_\gamma(x) - T_\gamma(y)))\|_2 + \|\nabla^2 f(x) - \nabla^2 f(y)\|_2(y - T_\gamma(y))\|_2 \leq 1 + 2\gamma \lambda_2 \|x - y\| + \lambda_3(\|x\|_2 + R)\|x - y\|_2 \leq (\gamma^{-1} + 2\lambda_2 + \lambda_3(\lambda_0 + R))\|x - y\|_2 = \lambda_2 \|x - y\|_2. \quad (B22)
\]
Above, (i) uses Proposition 2.3, (ii) uses (3d) and (B18)–(B20), and (iii) uses the assumption that \( \|y\|_2 \leq \lambda_0 \), without loss of generality. Indeed, if both \( \|x\|_2 \geq \lambda_0 \) and \( \|y\|_2 \geq \lambda_0 \), then \( f(x) = f(y) = \|\nabla f(x)\|_2 = \|\nabla f(y)\|_2 = \|\nabla^2 f(x)\| = \|\nabla^2 f(y)\| = 0 \) by (3), and thus (B22) still holds. This observation justifies our earlier restriction to the case where \( \|y\|_2 \leq \lambda_0 \).

Next, we show the strong convexity of \( F^\text{FB}_\gamma \) over long distances. Again, without loss of generality, we assume below
that \( \|y\|_2 \leq \lambda_0 \) and write

\[
\langle x - y, \nabla F_{\gamma}^{FB}(x) - \nabla F_{\gamma}^{FB}(y) \rangle \\
\quad \equiv \left\{ \frac{1}{\gamma} \langle x - y, (I_d - \gamma \nabla^2 f(x))(x - T_\gamma(x)) \rangle \right\} \\
\quad \quad - \left\{ \frac{1}{\gamma} \langle x - y, (I_d - \gamma \nabla^2 f(y))(y - T_\gamma(y)) \rangle \right\} + \frac{1}{\gamma} \left\{ \langle x - y, (\nabla^2 f(x) - \nabla^2 f(y))(y - T_\gamma(y)) \rangle \right\} \\
\quad \geq \frac{1}{\gamma} \langle x - y, (I_d - \gamma \nabla^2 f(x))(x - y) \rangle \\
\quad \quad - \frac{\|I_d - \gamma \nabla^2 f(x)\|}{\gamma} \|x - y\|_2 (\|T_\gamma(x)\|_2 + \|T_\gamma(y)\|_2) \\
\quad \quad - \|x - y\|_2 \cdot \|\nabla^2 f(x) - \nabla^2 f(y)\| \cdot \|y - T_\gamma(y)\|_2 \\
\quad \quad \geq \frac{1 - \gamma \lambda_2}{\gamma} \|x - y\|_2^2 - \frac{2R}{\gamma} \|x - y\|_2 \\
\quad \quad - \lambda_3 (\lambda_0 + R) \|x - y\|_2^2 \\
\quad \quad = \frac{1}{\gamma} \left( 1 - \gamma (\lambda_2 + \lambda_3 (\lambda_0 + R)) \right) \|x - y\|_2^2 \\
\quad \quad \geq \frac{1}{\gamma} \left( \frac{2R}{1 - 2\gamma (\lambda_2 + \lambda_3 (\lambda_0 + R))} \right) \\
\end{equation}

where (i) uses Proposition 2.3 and (ii) uses (3d), (B17), (B18), and (B20). When \( \|x - y\|_2 \) is sufficiently large, the first term in the last line above dominates the second term. In particular, we have

\[
\|x - y\|_2 \geq \rho_{\gamma}^{FB} := \frac{2R}{1 - 2\gamma (\lambda_2 + \lambda_3 (\lambda_0 + R))} \\
\implies \langle x - y, \nabla F_{\gamma}^{FB}(x) - \nabla F_{\gamma}^{FB}(y) \rangle \geq (\lambda_2 + \lambda_3 (\lambda_0 + R)) \|x - y\|_2^2. 
\]

In words, \( F_{\gamma}^{FB} \) behaves like a strongly convex function over long distances. It is not difficult to verify that Definition 1.2 is also valid for the FB envelope: To that end, one needs to combine Proposition 1 with the relation between the MY and FB envelopes in Proposition 2.2.

Lastly, we now verify that the requirement in Definition 1.1 is met for the FB envelope: Below, suppose that \( \gamma \in (0, 1/\lambda_2) \). On the one hand, Proposition 2.2 implies that

\[
F_{\gamma}^{FB} \geq F_{1/\gamma^2}^{MY}. 
\]

On the other hand, Proposition 1 implies that there exists \( F^0 \) such that

\[
F_{1/\gamma^2}^{MY} \geq F^0, 
\]

and \( \exp(-F^0) \) is integrable. By combining the preceding two observations, we find that the FB envelope satisfies Definition 1.1 when \( \gamma \in (0, 1/\lambda_2) \). This completes the proof of Proposition 3.

### Appendix C: Proof of Theorem 1

For every \( x \in \mathbb{R}^d \), let us define

\[
q(x) := \frac{e^{f(x) - 1_{K}(x)}}{\int_{\mathbb{R}^d} e^{f(z) - 1_{K}(z)} \, dz}, 
\]

where \( 1_K \) is the indicator function on the set \( K \). Note that the target distribution \( p \) in (6a) coincides with \( q \) in the special case where \( g = 1_K \). Let \( \text{dist}(x, K) \) denote the distance from \( x \) to the set \( K \). For \( \gamma > 0 \), we also define

\[
1_{K,\gamma}(x) := \frac{1}{2\gamma} \text{dist}(x, K)^2 = \frac{1}{2\gamma} \min_{z \in \mathbb{R}^d} \left\{ \frac{1}{2} \|x - z\|_2^2 \right\}, 
\]

to be the MY envelope of the indicator function on the set \( K \). We denote the corresponding probability distribution by

\[
q_{\gamma}(x) := \frac{e^{f(x) - 1_{K,\gamma}(x)}}{\int_{\mathbb{R}^d} e^{f(z) - 1_{K,\gamma}(z)} \, dz}, \quad x \in \mathbb{R}^d. 
\]

When \( \gamma \) is sufficiently small, we may intuitively regard \( q_{\gamma} \) as a proxy for \( q \). We also recall our earlier notation for the convenience of the reader:

\[
p(x) := \frac{e^{-F(x)}}{\int_{\mathbb{R}^d} e^{-F(z)} \, dz}, \\
p_{\gamma}(x) := \frac{e^{-F_{\gamma}(x)}}{\int_{\mathbb{R}^d} e^{-F_{\gamma}(z)} \, dz}, \quad x \in \mathbb{R}^d. 
\]

Above, the functions \( F \) and \( F_{\gamma}^{FB} \) were defined in (6b) and (FB), respectively. Our objective is to control the distance \( W_1(p_{\gamma}^{FB}, p) \). To begin, we use the triangle inequality to write that

\[
W_1(p_{\gamma}^{FB}, p) \leq W_1(p_{\gamma}^{FB}, q_{\gamma}) + W_1(q_{\gamma}, q) + W_1(q, p). 
\]

Each of the following three lemmas bounds one of the terms on the right-hand side above.

**Lemma 2** It holds that

\[
W_1(p_{\gamma}^{FB}, q_{\gamma}) \leq c_1 + c_3 \frac{I_2(y(1 - \gamma \lambda_2))}{d(K) + I_1(y)}, 
\]

where \( 1_{K} \) is the indicator function on the set \( K \). Note that the target distribution \( p \) in (6a) coincides with \( q \) in the special case where \( g = 1_{K} \). Let \( \text{dist}(x, K) \) denote the distance from \( x \) to the set \( K \). For \( \gamma > 0 \), we also define

\[
1_{K,\gamma}(x) := \frac{1}{2\gamma} \text{dist}(x, K)^2 = \frac{1}{2\gamma} \min_{z \in \mathbb{R}^d} \left\{ \frac{1}{2} \|x - z\|_2^2 \right\}, 
\]

to be the MY envelope of the indicator function on the set \( K \). We denote the corresponding probability distribution by

\[
q_{\gamma}(x) := \frac{e^{f(x) - 1_{K,\gamma}(x)}}{\int_{\mathbb{R}^d} e^{f(z) - 1_{K,\gamma}(z)} \, dz}, \quad x \in \mathbb{R}^d. 
\]

When \( \gamma \) is sufficiently small, we may intuitively regard \( q_{\gamma} \) as a proxy for \( q \). We also recall our earlier notation for the convenience of the reader:

\[
p(x) := \frac{e^{-F(x)}}{\int_{\mathbb{R}^d} e^{-F(z)} \, dz}, \\
p_{\gamma}(x) := \frac{e^{-F_{\gamma}(x)}}{\int_{\mathbb{R}^d} e^{-F_{\gamma}(z)} \, dz}, \quad x \in \mathbb{R}^d. 
\]

Above, the functions \( F \) and \( F_{\gamma}^{FB} \) were defined in (6b) and (FB), respectively. Our objective is to control the distance \( W_1(p_{\gamma}^{FB}, p) \). To begin, we use the triangle inequality to write that

\[
W_1(p_{\gamma}^{FB}, p) \leq W_1(p_{\gamma}^{FB}, q_{\gamma}) + W_1(q_{\gamma}, q) + W_1(q, p). 
\]

Each of the following three lemmas bounds one of the terms on the right-hand side above.

**Lemma 2** It holds that

\[
W_1(p_{\gamma}^{FB}, q_{\gamma}) \leq c_1 + c_3 \frac{I_2(y(1 - \gamma \lambda_2))}{d(K) + I_1(y)}, 
\]
where
\[ c_1 := \frac{e^{\max_{x \in K} f(x) - \gamma \lambda_2} \min \left\{ f(x) \right\} \int_k \|e^{(1-\gamma \lambda_2) f(x)}\|_2^2 \, dx - \int_k e^{-(1-\gamma \lambda_2) f(x)} \, dx}{e^{\gamma \lambda_2} \min_{x \in K} \int_k \|e^{f(x)}\|_2^2 \, dx - \int_k e^{f(x)} \, dx} - \max \left( \frac{e^{-f(x)} - \int_k \|e^{-f(x)}\|_2^2 \, dx}{e^{\gamma \lambda_2} \min_{x \in K} \int_k \|e^{f(x)}\|_2^2 \, dx - \int_k e^{f(x)} \, dx} \right), \]
\[ c_3 := e^{\max_{x \in K} f(x) - \min_{x \in K} g(x)}, \]
\[ I_1(\gamma) := \sum_{i=0}^{d-1} d_i(K) \cdot (2\pi \gamma)^{d-1}, \]
\[ I_2(\gamma) := \sum_{i=0}^{d-1} d_i(K) \cdot (2\pi \gamma)^{d-1} \left( \sqrt{\gamma (d-i+3)} + R \right), \]
where \( d_i(K) \) is the \( i \)-th intrinsic volume of \( K \) (Klain and Rota 1997). In particular, the \( d \)-th volume of \( K \) coincides with the standard volume of \( K \), i.e., \( d_d(K) = d(K) \).

**Lemma 3** It holds that
\[ W_1(q_{\gamma}, \gamma) \leq c_2 - R I_1(\gamma) + I_2(\gamma), \quad (C31) \]
where \( c_2 := e^{\max_{x \in K} f(x) - \min_{x \in K} f(x)} \).

**Lemma 4** It holds that
\[ W_1(q, p) \leq c_4 R, \quad (C32) \]
where
\[ c_4 := e^{\max_{x \in K} g(x) - \min_{x \in K} g(x)} - e^{\min_{x \in K} g(x) - \max_{x \in K} g(x)} \].

To keep the notation light, above we have suppressed the dependence of \( c_1 \) to \( c_4 \) on \( K, f, g \). With Lemmas 2–4 at hand, we revisit (C29) and write
\[ W_1(p_{\gamma}^{FB}, p) \leq c_1 + c_2 R I_1(\gamma) + c_2 I_2(\gamma) + c_3 I_2(\gamma/(1 - \gamma \lambda_2)) - c_4 R, \quad (C34) \]
which completes the proof of Proposition 1.

**Appendix D: Proof of Lemma 2**

In order to upper bound the distance \( W_1(q_{\gamma}, p_{\gamma}^{FB}) \), we will use the following result which bounds the \( W_1 \) distance under small perturbations.

\[ \textbf{Lemma 5} \quad \text{For constants } \alpha \geq \beta \text{ and } \beta' \leq 1, \text{ consider } h_1 : \mathbb{R}^d \to \mathbb{R} \text{ and } h_2 : \mathbb{R}^d \to \mathbb{R} \text{ that satisfy} \]
\[ \beta' \cdot h_1(x) + \beta \leq h_2(x) \leq h_1(x) + \alpha, \quad (D35) \]
for every \( x \in \mathbb{R}^d \). Then it holds that
\[ W_1\left( h_1(x), h_2(x) \right) \leq \frac{e^{\beta' \cdot h_1(x)} - e^{\beta \cdot h_2(x)}}{e^{\beta - e^{-h_1(x)}}} \]
\[ - \max \left( \frac{e^{\beta - e^{-h_1(x)}} \int_k \|e^{h_1(x)}\|_2^2 \, dx - \int_k e^{-h_1(x)} \, dx}{e^{\beta - e^{-h_1(x)}} \int_k \|e^{-h_1(x)}\|_2^2 \, dx - \int_k e^{-h_1(x)} \, dx} \right). \quad (D36) \]

As a sanity check, note that the right-hand side of (D36) is always nonnegative because, by assumption, \( \alpha \geq \beta \) and \( \beta' \leq 1 \). Moreover, if we set \( \alpha = \beta = 0 \) and \( \beta' = 1 \), the right-hand side of (D36) reduces to zero, as expected.

Let us now recall from (C27) and (6a) that \( q_{\gamma} \propto \exp(-f - 1_{K_{1,\gamma}}) \) and \( p_{\gamma}^{FB} \propto \exp(-F_{\gamma}^{FB}) \), respectively. Our plan is to invoke Lemma 5 with the choice of \( h_1 = f + 1_{K,\gamma} \) and \( h_2 = F_{\gamma}^{FB} \). In turn, this plan necessitates that we verify (D35) for the choice of \( h_1 = f + 1_{K,\gamma} \) and \( h_2 = F_{\gamma}^{FB} \). We begin by relating the two functions as
\[ F_{\gamma}^{FB}(x) \leq F_{\gamma}^{MY}(x) \]
\[ \Leftrightarrow f(x) + g(x) \leq f(x) + g(x) \]
\[ \leq f(x) + \min_{z \in K} \left\{ g(z) + \frac{1}{2\gamma} \| x - z \|_2^2 \right\} \]
\[ \leq \max_{z \in K} g(z) + f(x) + \frac{1}{2\gamma} \min_{z \in K} \| x - z \|_2^2 \]
\[ \Leftrightarrow \max_{z \in K} g(z) + f(x) + 1_{K,\gamma}(x), \quad (C37) \]
where (i) uses Proposition 2.2, (ii) uses (MY), (iii) uses (8) and the fact that \( g \) is by (5) infinite outside of \( K \), and (iv) uses (C26). Above, note that \( \max_{z \in K} g(z) \) is finite by the construction of \( g \); see (5). In the other direction, we similarly write
\[ F_{\gamma}^{FB}(x) \]
\[ \geq F_{\gamma}^{MY}(x) \]
\[ \Leftrightarrow f(x) + g(x) \geq f(x) + g(x) \]
\[ \geq f(x) + \min_{z \in K} \left\{ g(z) + \frac{1 - \gamma \lambda_2}{2\gamma} \| x - z \|_2^2 \right\} \]
\[ \geq - \max_{z \in K} g(z) + f(x) + \frac{1 - \gamma \lambda_2}{2\gamma} \min_{z \in K} \| x - z \|_2^2 \]
\[ \Leftrightarrow - \max_{z \in K} g(z) + \gamma \lambda_2 f(x) + (1 - \gamma \lambda_2) f(x) \]
\[ + (1 - \gamma \lambda_2) 1_{K,\gamma}(x) \]
\[
\begin{align*}
\beta'(f(x) + 1)_{K_{\gamma}}(x) + \beta & \leq F^\text{FB}_{\gamma}(x) \\
\leq f(x) + 1_{K_{\gamma}}(x) + \alpha, & \quad \text{(C39)}
\end{align*}
\]

where
\[
\alpha = \max_{z \in K} g(z), \quad \beta = \gamma \lambda_2 \min_{z \in K} f(z) - \max_{z \in K} g(z),
\]
\[
\beta' = 1 - \gamma \lambda_2.
\]

With (C39) at hand, we can invoke Lemma 5 to find that
\[
W_1(q_{\gamma}, p_{\gamma}^\text{FB})
\]
\[
\leq \begin{aligned}
\int e^{-f(x) - \gamma \lambda_2 f(x)} \, dx \\
\geq - \max_{z \in K} g(z) + \gamma \lambda_2 \min_{z \in K} f(z) + (1 - \gamma \lambda_2) f(x)
\end{aligned}
\]
\[
\geq (1 - \gamma \lambda_2) 1_{K_{\gamma}}(x), & \quad \text{(C38)}
\]

where (i) uses Proposition 2.2 and (ii) uses (C26). Above, \( \min_{z} f(z) \) is finite by the construction of \( f \); see (3). We now summarize (C37) and (C38); (D35) is satisfied for our choice of \( h_1 = f + 1_{K_{\gamma}} \) and \( h_2 = F^\text{FB}_{\gamma} \), i.e.,

\[
\beta'(f(x) + 1)_{K_{\gamma}}(x) + \beta \leq F^\text{FB}_{\gamma}(x)
\]

\[
\leq f(x) + 1_{K_{\gamma}}(x) + \alpha, & \quad \text{(C39)}
\]

Lemma 6 It holds that
\[
\int e^{-f(x) - \gamma \lambda_2 f(x) \cdot \text{dist}(x, K^c)^2 / \gamma^2} \, dx
\]
\[
= \sum_{i=0}^{d-1} d_i(K) \cdot (2 \pi \gamma)^{d-i} I_1(\gamma), & \quad \text{(C41a)}
\]

\[
\int e^{-f(x) - \gamma \lambda_2 f(x) \cdot \text{dist}(x, K^c)^2 / \gamma^2} \, dx
\]
\[
= \sum_{i=0}^{d} d_i(K) \cdot (2 \pi \gamma)^{d-i} = d(K) + I_1(\gamma), & \quad \text{(C41b)}
\]

\[
\int_{K^c} \|x\|e^{-\text{dist}(x, K^c)^2 / \gamma^2} \, dx \leq I_2(\gamma)
\]
\[
:= \sum_{i=0}^{d} d_i(K) \cdot (2 \pi \gamma)^{d-i} \gamma(\sqrt{d - i + 3} + R), & \quad \text{(C41c)}
\]

where \( K^c = \mathbb{R}^d \setminus K \) is the complement of the set \( K \) and \( d_i(K) \) is the i-th intrinsic volume of \( K \) (Klain and Rota 1997). In particular, the d-th volume of \( K \) coincides with the standard volume of \( K \), i.e., \( d_d(K) = d(K) \).

As a sanity check for Lemma 6, note that \( I_1(\gamma) \) and \( I_2(\gamma) \) both vanish in the limit of \( \gamma \to 0 \), as expected. To control the first fraction in the last line of (C40), we use Lemma 6 to write that
\[
\int \|x\|^2 e^{-\gamma \lambda_2 f(x) \cdot \text{dist}(x, K)^2 / \gamma^2} \, dx
\]
\[
\leq \begin{aligned}
\int \|x\|^2 e^{-\gamma \lambda_2 f(x)} \, dx \\
\leq \gamma \lambda_2 \min_{z \in K} f(z)
\end{aligned}
\]

\[
\int \|x\|^2 e^{-f(x) - \gamma \lambda_2 f(x) \cdot \text{dist}(x, K)^2 / \gamma^2} \, dx
\]

where (i) breaks the numerator into two integrals on \( K \) and its complement, and uses the fact that \( \text{dist}(x, K) = 0 \) when \( x \in K \). Above, (ii) uses Lemma 6. For the second fraction in the last line of (C40), we note that
\[
\int \|x\|^2 e^{-f(x) - \gamma \lambda_2 f(x) \cdot \text{dist}(x, K)^2 / \gamma^2} \, dx
\]

\[
\int \|x\| e^{-f(x) - \gamma \lambda_2 f(x) \cdot \text{dist}(x, K)^2 / \gamma^2} \, dx
\]

\[
= \sum_{i=0}^{d} d_i(K) \cdot (2 \pi \gamma)^{d-i} I_1(\gamma), & \quad \text{(C42)}
\]
\[ W_{1}(q, p_{FB}) \leq \frac{e^{\gamma_{\lambda} \min_{\nu \in \mathcal{F}_{2}} f(\nu)} - e^{\gamma_{\lambda} \min_{\nu \in \mathcal{F}_{2}} f(\nu)} \| x \|_{2}^{2}}{\max \left( e^{\gamma_{\lambda} \min_{\nu \in \mathcal{F}_{2}} f(\nu)} - e^{\gamma_{\lambda} \min_{\nu \in \mathcal{F}_{2}} f(\nu)} \| x \|_{2}^{2}, e^{-\gamma_{\lambda} \min_{\nu \in \mathcal{F}_{2}} f(\nu)} \right) \| x \|_{2}^{2}} + e^{\gamma_{\lambda} \min_{\nu \in \mathcal{F}_{2}} f(\nu)} \| x \|_{2}^{2} \cdot i_{2}(y) \cdot i_{1}(y), \] (C45)

which completes the proof of Lemma 2.

**Appendix E: Proof of Lemma 5**

Let us first recall Theorem 6.15 in Villani (2009), which we will use in the proof to link the \( W_{1} \) and TV distances of two rapidly decaying distributions.

**Theorem 3** Any pair \((\tau_{1}, \tau_{2})\) of probability distributions satisfies

\[ W_{1}(\tau_{1}, \tau_{2}) \leq \int \| x \|_{2} \cdot |\tau_{1}(x)| - \tau_{2}(x)| \, dx. \] (E46)

We apply Theorem 3 to bound the \( W_{1} \) distance of interest as

\[ W_{1}(\frac{e^{-h_{1}}}{f_{e^{-h_{1}}}}, \frac{e^{-h_{2}}}{f_{e^{-h_{2}}}}) \leq \int_{\mathbb{R}^{d}} \| x \|_{2} \left| \frac{e^{-h_{1}(x)}}{f_{e^{-h_{1}(x)}}} - \frac{e^{-h_{2}(x)}}{f_{e^{-h_{2}(x)}}} \right| \, dx. \] (C47)

To bound the integral on the right-hand side above, we first use the assumed relationship between \( h_{1} \) and \( h_{2} \) to write that

\[ \int \frac{e^{-h_{1}(x)}}{f_{e^{-h_{1}(x)}}} \, dz - \int \frac{e^{-h_{2}(x)}}{f_{e^{-h_{2}(x)}}} \, dz \]

\[ \leq \int \frac{e^{-h_{1}(x)} - e^{-h_{2}(x)}}{f_{e^{-h_{1}(x)}}} \, dz - \int \frac{e^{-h_{1}(x) - \alpha}}{f_{e^{-h_{1}(x) - \alpha}}} \, dz \]

\[ = \int \frac{e^{-h_{1}(x)}}{f_{e^{-h_{1}(x) - \alpha}}} \, dz - \int \frac{e^{-h_{1}(x) - \alpha}}{f_{e^{-h_{1}(x) - \alpha}}} \, dz \]

\[ = \Delta(x), \] (C48)

where (i) uses (D35). In the other direction, we can again use the relation between \( h_{1} \) and \( h_{2} \) as

\[ \int \frac{e^{-h_{1}(x)}}{f_{e^{-h_{1}(x)}}} \, dz - \int \frac{e^{-h_{2}(x)}}{f_{e^{-h_{2}(x)}}} \, dz \]

\[ \geq \int \frac{e^{-h_{1}(x) - \beta}}{f_{e^{-h_{1}(x) - \beta}}} \, dz - \int \frac{e^{-h_{2}(x) - \alpha}}{f_{e^{-h_{2}(x) - \alpha}}} \, dz \]

\[ \geq \int \frac{e^{-h_{1}(x)}}{f_{e^{-h_{1}(x) - \beta}}} \, dz - \int \frac{e^{-h_{1}(x) - \beta}}{f_{e^{-h_{1}(x) - \beta}}} \, dz \]

\[ = \Delta(x), \] (C49)

where (i) uses (D35) and (ii) uses the assumptions that \( \alpha \geq \beta \) and \( \beta' \leq 1 \). By combining (C48) and (C49), we find that

\[ \int \frac{e^{-h_{1}(x)}}{f_{e^{-h_{1}(x)}}} \, dz - \int \frac{e^{-h_{1}(x)}}{f_{e^{-h_{1}(x)}}} \, dz \]

\[ \geq \int \frac{e^{-h_{1}(x)}}{f_{e^{-h_{1}(x) - \beta}}} \, dz - \int \frac{e^{-h_{1}(x) - \beta}}{f_{e^{-h_{1}(x) - \beta}}} \, dz \]

\[ \Delta(x) \leq 0, \] (C50)

where (i) uses (C48), (C49), and the following observation:

\[ a \leq b \leq c \text{ and } a \leq 0 \text{ imply that} \]

\[ |b| \leq \begin{cases} -a & c \leq 0, \\ -a + c & a \geq 0. \end{cases} \] (C51)

It immediately follows from (C50) that

\[ \int \| x \|_{2} \left| \frac{e^{-h_{1}(x)}}{f_{e^{-h_{1}(x)}}} \, dz - \frac{e^{-h_{2}(x)}}{f_{e^{-h_{2}(x)}}} \, dz \right| \, dx \]

\[ \leq \int_{\Delta(x) \leq 0} \| x \|_{2} e^{-\beta h_{1}(x) + \alpha - \beta} \, dx \]

\[ - \int \Delta(x) \leq 0 \| x \|_{2} e^{-h_{2}(x)} \, dx \]

\[ + \int_{\Delta(x) \geq 0} \| x \|_{2} e^{-h_{1}(x) + \alpha - \beta} \, dx \]

\[ - \int \Delta(x) \geq 0 \| x \|_{2} e^{-h_{2}(x)} \, dx \]

\[ \Delta(x) \leq 0, \] (C52)
\[
\begin{align*}
& (i) \quad e^{a-b} \int \frac{\|x\|^2 e^{-\beta h_1(x)}}{e^{-h_1(z)}} \, dz \\
& \leq \frac{e^{b-a} \int_{\Delta(x) \leq 0} \|x\|^2 e^{-h_1(x)}}{e^{-h_1(z)}} \, dz \\
& \quad - \frac{e^{b-a} \int_{\Delta(x) > 0} \|x\|^2 e^{-h_1(x)}}{e^{-h_1(z)}} \, dz \\
& \qquad - \frac{e^{b-a} \int \|x\|^2 e^{-h_1(x)}}{e^{-h_1(z)}} \, dz \\
& \leq \frac{e^{a-b} \int \|x\|^2 e^{-\beta h_1(x)}}{e^{-h_1(z)}} \, dz \\
& \quad - \frac{e^{b-a} \int \|x\|^2 e^{-h_1(x)}}{e^{-h_1(z)}} \, dz \\
& \quad - \max(\int e^{-h_1(z)} \, dz, \frac{e^{-\beta h_1(x)}}{e^{-h_1(z)}} \, dz),
\end{align*}
\]

completes the proof of Lemma 5.

**Appendix F: Proof of Lemma 6**

Let \( \iota(A) = 1 \) if the claim \( A \) is true and \( \iota(A) = 0 \) otherwise. Using the fact that
\[
e^{-\frac{dist(x, K)^2}{2\gamma}} = \frac{1}{\gamma} \int_0^\infty t e^{-\frac{t^2}{\gamma}} \, dt,
\]
we can rewrite the left-hand side of (C41b) as
\[
\int_{R^d} e^{-\frac{dist(x, K)^2}{2\gamma}} \, dx \\
= \frac{1}{\gamma} \int_{R^d} \int_0^\infty t e^{-\frac{t^2}{\gamma}} \, dt \, dx \\
= \frac{1}{\gamma} \int_{R^d} \int_0^\infty \iota(dist(x, K) \leq t) t e^{-\frac{t^2}{\gamma}} \, dt \, dx \\
= \frac{1}{\gamma} \int_0^\infty \left( \int_{R^d} \iota(dist(x, K) \leq t) \, dx \right) \cdot t e^{-\frac{t^2}{\gamma}} \, dt \\
= \frac{1}{\gamma} \int_0^\infty d(dist(x, K) \leq t) t e^{-\frac{t^2}{\gamma}} \, dt.
\]

Note that \( \{x : dist(x, K) \leq t\} \) is the tube of radius \( t \) around the set \( K \). We can represent this set more compactly as \( K + B(0, t) \), where the addition is in the Minkowski sense. With this in mind, we rewrite the last line above as
\[
\int_{R^d} e^{-\frac{dist(x, K)^2}{2\gamma}} \, dx \\
= \frac{1}{\gamma} \int_0^\infty d(K + B(0, t)) \cdot t e^{-\frac{t^2}{\gamma}} \, dt.
\]

We can now use the Steiner’s formula (Klain and Rota 1997) to express the volume above as
\[
d(K + B(0, t)) = \sum_{i=0}^d \kappa_i d_{d-i}(K),
\]
where \( \kappa_i \) is the volume of the unit ball in \( R^i \) and \( d_i(K) \) is the \( i \)-th intrinsic volume of \( K \) (Klain and Rota 1997). In particular, the \( d \)-th intrinsic volume of \( K \) coincides with its standard volume, i.e., \( d_d(K) = d(K) \).

We thus established (C41b). Above, \( (ii) \) also uses the identity
\[
\int_0^\infty t^i e^{-\frac{t^2}{\gamma}} \, dt = \gamma (2\gamma)^{i/2} \Gamma \left( \frac{i+1}{2} \right).
\]

To prove (C41a), we write that...
\[
\int_{K^c} e^{-\frac{\text{dist}(x, K)^2}{2\gamma}} \, dx
\]

\[
= \int_{\mathbb{R}^d} e^{-\frac{\text{dist}(x, K)^2}{2\gamma}} \, dx - \int_K e^{-\frac{\text{dist}(x, K)^2}{2\gamma}} \, dx
\]

\[
\overset{(i)}{=} \text{d}(K) + I_1(\gamma) - \int_K \, dx
\]

\[
= I_1(\gamma), \quad (\text{F59})
\]

where (i) uses (C41b) and the fact that dist(x, K) = 0 when \(x \in K\). To prove (C41c), we use the fact that \(K \subset B(0, R)\) by Assumption 1.1, which implies that

\[
\text{dist}(x, K) \geq \|x\|_2 - R. \quad (\text{F60})
\]

In turn, we use (F60) to write that

\[
\int_{K^c} \|x\|_2^2 e^{-\frac{\text{dist}(x, K)^2}{2\gamma}} \, dx
\]

\[
\overset{(i)}{\leq} \int_{\text{dist}(x, K) > 0} (\text{dist}(x, K) + R) e^{-\frac{\text{dist}(x, K)^2}{2\gamma}} \, dx
\]

\[
\overset{(ii)}{=} \int_{\text{dist}(x, K) > 0} \left( \int_{\text{dist}(x, K)}^\infty \frac{t}{\Gamma(1+i/2)} e^{-\frac{t^2}{\gamma}} \, dt \right) \, dx
\]

\[
= \frac{1}{\gamma} \int_{\mathbb{R}^d} \left( \int_0^\infty 0 < \text{dist}(x, K) \leq t \right) \left( t + R \right) e^{-\frac{t^2}{2\gamma}} \, dt \, dx
\]

\[
\overset{(iii)}{\leq} \frac{1}{\gamma} \int_{\mathbb{R}^d} \left( \int_0^\infty 0 < \text{dist}(x, K) \leq t \right) \left( t + R \right) e^{-\frac{t^2}{2\gamma}} \, dt \, dx
\]

\[
= \frac{1}{\gamma} \int_0^\infty \left( \int_{\mathbb{R}^d} 0 < \text{dist}(x, K) \leq t \right) \left( t + R \right) e^{-\frac{t^2}{2\gamma}} \, dt
\]

\[
= \frac{1}{\gamma} \int_0^\infty \left( \text{d}(K + B(0, t)) - \text{d}(K) \right) \left( t + R \right) e^{-\frac{t^2}{2\gamma}} \, dt, \quad (\text{F61})
\]

where (i) uses (F60) and the fact that \(K^c = \{x : \text{dist}(x, K) > 0\}\) and (ii) uses (F54). We can again use the Steiner’s formula to calculate the volume of the tube of radius t around K, which appears in the last line above. Substituting from (F56) into the last line above, we find that

\[
\int_{K^c} \|x\|_2^2 e^{-\frac{\text{dist}(x, K)^2}{2\gamma}} \, dx
\]

\[
\overset{(i)}{\leq} \frac{1}{\gamma} \int_0^\infty \sum_{i=1}^d t' \kappa_i \text{d}_{d-i}(K) \cdot (t + R) t e^{-\frac{t^2}{2\gamma}} \, dt
\]

\[
= \frac{1}{\gamma} \sum_{i=1}^d \kappa_i \text{d}_{d-i}(K) \left( t^{i+2} + R t^{i+1} \right) e^{-\frac{t^2}{2\gamma}} \, dt
\]

\[
= \sum_{i=1}^d \kappa_i \text{d}_{d-i}(K) (2\gamma)^{i+1} \Gamma \left( \frac{i + 3}{2} \right)
\]

\[
+ \sum_{i=1}^d \kappa_i \text{d}_{d-i}(K) R (2\gamma)^{i} \frac{1}{2} \Gamma \left( \frac{i + 1}{2} \right)
\]

\[
\overset{(ii)}{=} \sum_{i=1}^d \frac{\pi^i}{\Gamma(1+i/2)} \text{d}_{d-i}(K) (2\gamma)^{i+1} \Gamma \left( \frac{i + 3}{2} \right)
\]

\[
+ \sum_{i=1}^d \frac{\pi^i}{\Gamma(1+i/2)} \text{d}_{d-i}(K) R (2\gamma)^{i} \frac{1}{2} \Gamma \left( \frac{i + 1}{2} \right)
\]

\[
= \sum_{i=1}^d \text{d}_{d-i}(K) (2\gamma)^{i} \left( \sqrt{2\gamma} \cdot \Gamma \left( \frac{i + 3}{2} \right) + R \right)
\]

\[
= \sum_{i=1}^d \text{d}_{d-i}(K) (2\gamma)^{i} \left( \sqrt{2\gamma} \cdot \sqrt{i + 3} + R \right)
\]

\[
= \sum_{i=0}^d \text{d_i}(K) (2\gamma)^{\frac{d-i}{2}} \left( \sqrt{2\gamma} (d - i + 3) + R \right)
\]

\[
= I_2(\gamma), \quad (\text{F62})
\]

where (i) uses (F56), (F61), and the fact that \(d_d(K) = d(K)\). Also, (ii) uses (F58) and the fact that \(\kappa_i = \pi^i / \Gamma(1+i/2)\). Above, (iii) is an application of the Gautschi’s inequality. This establishes (C41c) and completes the proof of Lemma 6.

### Appendix G: Proof of Lemma 3

Recall from (C25) that \(q \propto \exp(-f - 1_K(x))\), where \(1_K\) denotes the indicator function on the set \(K\). Recall also from (C27) that \(q_{\gamma} \propto \exp(-f - 1_{K_{\gamma}}(x))\), where \(1_{K_{\gamma}}(x) = \frac{1}{\gamma} \text{dist}(x, K)^2\) is the MY envelope of \(1_K\); see (C26). We will repeatedly use these two distributions in the proof. To begin, let us invoke Theorem 3 to write that

\[
W_1(q_{\gamma}, q) \leq \int \|x\|_2 \cdot |q_{\gamma}(x) - q(x)| \, dx
\]

\[
\overset{(i)}{=} \int_{K} \|x\|_2 \cdot |q_{\gamma}(x) - q(x)| \, dx
\]

\[
+ \int_{K^c} \|x\|_2 \cdot |q_{\gamma}(x) - q(x)| \, dx, \quad (\text{G63})
\]

where the first line above uses Theorem 3. For the first integral on the right-hand side of (i) above, we use the fact that \(K \subset \mathbb{R}^d\) and the convexity of \(\text{dist}^2(x, K)\) to write

\[
\int_{K} \|x\|_2 \cdot |q_{\gamma}(x) - q(x)| \, dx
\]

\[
\overset{(ii)}{=} \int_{K} \|x\|_2 \cdot |q_{\gamma}(x) - q(x)| \, dx
\]

\[
+ \int_{K^c} \|x\|_2 \cdot |q_{\gamma}(x) - q(x)| \, dx, \quad (\text{G63})
\]
B(0, R) in Assumption 1.1 to write that

\[
\int_K \|x\|_2 \cdot |q_y(x) - q(x)| \, dx \\
\leq R \int_K |q_y(x) - q(x)| \, dx
\]

\[
\leq R \int_K \left| e^{-f(x)} \left( \frac{e^{-f(x)}}{\int_{\mathbb{R}^d} e^{-f(z)} \, dz} - \frac{e^{-f(x)}}{\int_K e^{-f(z)} \, dz} \right) \right| \, dx
\]

\[
= R \int_K e^{-f(x)} \, dx
\]

\[
\cdot \left( \frac{1}{\int_K e^{-f(z)} \, dz} - \frac{1}{\int_{\mathbb{R}^d} e^{-f(z)} \, dz} \right),
\]

where \((i)\) uses Assumption 1.1, \((ii)\) uses the fact that \(1_K(x) = 1_K, y(x) = 0\) when \(x \in K\), and \((iii)\) uses the fact that

\[
\int_{\mathbb{R}^d} e^{-f(z) - \frac{\text{dist}(z, K)^2}{2\gamma}} \, dz \geq \int_K e^{-f(z) - \frac{\text{dist}(z, K)^2}{2\gamma}} \, dz
\]

\[
= \int_K e^{-f(z)} \, dz.
\]

We continue to simplify the last line of (G64) as

\[
\int_K \|x\|_2 \cdot |q_y(x) - q(x)| \, dx
\]

\[
\leq R \left( 1 - \frac{\int_K e^{-f(x)} \, dx}{\int_{\mathbb{R}^d} e^{-f(x) - \frac{\text{dist}(z, K)^2}{2\gamma}} \, dx} \right)
\]

\[
= R \cdot \int_{\mathbb{R}^d} e^{-f(x) - \frac{\text{dist}(z, K)^2}{2\gamma}} \, dz - \int_K e^{-f(x)} \, dx
\]

\[
= R \cdot \int_K e^{-f(x)} - \int_{\mathbb{R}^d} e^{-f(x) - \frac{\text{dist}(z, K)^2}{2\gamma}} \, dx
\]

\[
\leq R e^{\max g(x)} \cdot \frac{\int_K e^{-\frac{\text{dist}(z, K)^2}{2\gamma}} \, dx}{\int_{\mathbb{R}^d} e^{-\frac{\text{dist}(z, K)^2}{2\gamma}} \, dx}
\]

\[
= \frac{R e^{\max g(x)} \cdot \text{I}_1(y)}{d(K) + I_1(y)},
\]

where \((i)\) uses (G64) and \((ii)\) uses Lemma 6. For the second integral on the right-hand side of \((i)\) in (G63), we can again use the definitions of \(q\) and \(q_y\) to write that

\[
\int_{K^c} \|x\|_2 \cdot |q_y(x) - q(x)| \, dx
\]

\[
\leq \int_{K^c} \|x\|_2 \cdot q_y(x) \, dx
\]

\[
\leq \int_{K^c} \int_{\mathbb{R}^d} e^{-f(x) - \frac{\text{dist}(z, K)^2}{2\gamma}} \, dx
\]

\[
\leq \frac{e^{\max g(x)} \cdot \text{I}_2(y)}{d(K) + I_1(y)},
\]

where \((i)\) uses the fact that \(q\) is supported on \(K\) and \((ii)\) uses Lemma 6. With (G65) and (G66) at hand, we revisit (G63) and write that

\[
W_1(q_y, q) \leq e^{\max g(x) - \min g(x)} \frac{RI_1(y) + I_2(y)}{d(K) + I_1(y)}.
\]

This completes the proof of Lemma 3.

**Appendix H: Proof of Lemma 4**

Recall from (C25) and (6a) that \(q \propto \exp(-f - 1_K)\) and \(p \propto \exp(-F)\), respectively. The function \(F\) was defined in (6b). In order to bound \(W_1(q, p)\), our plan is to invoke Lemma 5 with \(h_1 = f + 1_K\) and \(h_2 = F\). As before, this means that we need to verify (D35) for the choice of \(h_1 = f + 1_K\) and \(h_2 = F\). We begin by relating these two functions together as

\[
F(x) = f(x) + g(x)
\]

\[
\leq \max_{z \in K} g(z) + f(x) + 1_K(x),
\]

where \((i)\) uses (6b), and \((ii)\) uses the fact that \(F\) and \(1_K\) both take infinity outside of the set \(K\). In the other direction, we write that

\[
F(x) = f(x) + g(x) + 1_K(x)
\]

\[
\geq \min_{z \in K} g(z) + f(x) + 1_K(x).
\]

To summarize, for our choice of \(h_1 = f + 1_K\) and \(h_2 = F\), (D35) is satisfied with

\[
\beta' (f(x) + 1_K(x)) + \beta \leq F(x)
\]

\[
\leq (f(x) + 1_K(x)) + \alpha,
\]

where

\[
\alpha = \max_{z \in K} g(z), \quad \beta = \min_{z \in K} g(z), \quad \beta' = 1.
\]

We can now invoke Lemma 5 to find that

\[
W_1(q, p) = W_1 \left( \frac{e^{-f} - 1_K}{\int e^{-f} - 1_K}, \frac{e^{-F}}{\int e^{-F}} \right)
\]

\[
\leq Z(K) \frac{\int \|x\|_2 e^{-f(x) - 1_K(x)} \, dx}{\int e^{-f(x) - 1_K(x)} \, dx}.
\]
where
\[ Z(K) = e^{\max_{x \in K} g(x)} - e^{\min_{x \in K} g(x)} \]

Above, (i) uses Lemma 5 and (ii) uses the fact that \( K \subseteq B(0, R) \) by Assumption 1.1. This completes the proof of Lemma 4.

### Appendix I: Proof of Lemma 1

Consider the stochastic differential equation
\[ d x_t = -\nabla F_{p_{\lambda y}}(x_t) + \sqrt{2} d B_t, \quad x_0 \sim p_{\lambda y}, \tag{I71} \]

where \( \{B_t\}_{t \geq 0} \) is the standard Brownian motion. Above, note that the initial probability measure is \( p_{\lambda y} \propto \exp(-F_{p_{\lambda y}}) \). From Definition 1.3, recall that \( F_{p_{\lambda y}} \) is \( \lambda_{\lambda y} \)-smooth and is, moreover, coercive. To be concrete, the latter means that \( F_{p_{\lambda y}} \) satisfies
\[ \lim_{||x||_2 \to \infty} F_{p_{\lambda y}}(x) = \infty. \]

Therefore Theorem 3.4 in Bou-Rabee and Eberle (2020) ensures that \( p_{\lambda y} \propto \exp(-F_{p_{\lambda y}}) \) is the invariant probability measure of (I71). In particular, because \( x_0 \sim p_{\lambda y} \) in (I71), it holds that
\[ x_t \sim p_{\lambda y}, \quad t \geq 0. \tag{I72} \]

Let \( \{P_k\}_{k \geq 0} \) denote the Markov transition kernel associated with the Markov chain \( \{x_{H_k}\}_{k \geq 0} \), where
\[ H_k := \sum_{i=1}^{k-1} h_i, \tag{I73} \]

is the elapsed time since initialization. Using this transition kernel, we can rewrite (I72) as
\[ p_{\lambda y} = p_{\lambda y} P_k, \quad k \geq 0. \tag{I74} \]

Finally, we can write the quantity of interest on the left-hand side of (26) as
\[ W_1(p_{\lambda y} Q_k, p_{\lambda y}) = W_1(p_{\lambda y} Q_k, p_{\lambda y} P_k), \tag{I75} \]

where we used (I74). In the remainder of the proof, we will upper bound the right-hand side above, which can be thought of as the discretization error associated with (I71). To begin, recall from (I72) that \( x_{H_k} \sim p_{\lambda y} \), and note that
\[
\begin{aligned}
x_{H_{k+1}} &= x_{H_k} - \int_{H_k}^{H_{k+1}} \nabla F_{p_{\lambda y}}(x_H) \, dt \\
&\quad + \sqrt{2} \int_{H_k}^{H_{k+1}} d B_t \sim p_{\lambda y} Q_k
\end{aligned}
\]

by construction. Above, (i) uses (20) and (ii) uses (73), i.e., the fact that \( H_{k+1} - H_k = h_k \). Likewise, note that
\[
\begin{aligned}
x_{H_{k+1}} &= x_{H_k} - \int_{H_k}^{H_{k+1}} \nabla F_{p_{\lambda y}}(x_H) \, dt \\
&\quad + \sqrt{2} \int_{H_k}^{H_{k+1}} d B_t \sim p_{\lambda y} P_k
\end{aligned}
\]

where (i) uses (I71) and (I74). In the second identity above, both sides of \( \text{dist.} \) have the same distribution and recall that \( \zeta_{k+1} \sim \text{normal}(0, I_d) \). In view of (I76) and (I77), we revisit (I75) and write that
\[ W_1(p_{\lambda y} Q_k, p_{\lambda y}) \]

where (i) uses (I71), (ii) uses (13), (176), and (177), (iii) uses (176) and (177), and (iii) uses Definition 1.3, i.e., the fact that \( F_{p_{\lambda y}} \) is \( \lambda_{\lambda y} \)-smooth. To estimate the expectation in the last line above, we write that
\[ E[\|x_t - x_{H_k}\|_2] \]

\[
\begin{aligned}
&\leq E \left\| \int_{H_k}^{H_{k+1}} \nabla F_{p_{\lambda y}}(x_H) \, dt \right\|_2 \\
&\quad + 2 E \left\| \int_{H_k}^{H_{k+1}} d B_t \right\|_2 \\
&\leq \lambda_{\lambda y} \int_{H_k}^{H_{k+1}} E[\|x_t - x_{H_k}\|_2^2] \, dt, \tag{I78}
\end{aligned}
\]

where (i) uses (I75), (ii) uses (13), (176), and (177), (iii) uses (176) and (177), and (iii) uses Definition 1.3, i.e., the fact that \( F_{p_{\lambda y}} \) is \( \lambda_{\lambda y} \)-smooth. To estimate the expectation in the last line above, we write that
\[
E[\|x_t - x_{H_k}\|_2] \]

\[
\begin{aligned}
&\leq E \left\| \int_{H_k}^{t} d x_s \right\|_2 \\
&\quad + 2 E \left\| \int_{H_k}^{t} \nabla F_{p_{\lambda y}}(x_s) \, ds + \sqrt{2} \int_{H_k}^{t} d B_s \right\|_2 \\
&\leq E \left\| \int_{H_k}^{t} \nabla F_{p_{\lambda y}}(x_s) \, ds \right\|_2 + \sqrt{2} E \left\| \int_{H_k}^{t} d B_s \right\|_2
\end{aligned}
\]
where (i) uses (I71). To estimate the first expectation in the last line of (I79), note that

\[
E_{x \sim p_{\gamma_k}} \| \nabla F_{\gamma_k}(x_s) \|_2 \leq h_k \sqrt{\mathbb{E}_1 \| \nabla F_{\gamma_k}(x_s) \|_2^2}.
\]

We now plug in the bounds in (I80) and (I82) back into (I79) to obtain that

\[
E_{x \sim p_{\gamma_k}} \| \nabla F_{\gamma_k}(x_s) \|_2 \leq \sqrt{h_k d} \left( \sqrt{h_k \lambda_{\gamma_k}} + \sqrt{2} \right),
\]

where (i) is simply (I79) and (ii) uses (I80) and (I82). By substituting the above bound back into (I78), we arrive at

\[
W_1(p_{\gamma_k} \mathcal{Q}_k, p_{\gamma_k}) \leq \lambda_{\gamma_k} \int_{H_k} E_{x \sim p_{\gamma_k}} \| x_t - x_{H_k} \|_2 dt
\]

\[
\leq \lambda_{\gamma_k} (H_k + 1 - H_k) \sqrt{h_k d} \cdot \left( \sqrt{h_k \lambda_{\gamma_k}} + \sqrt{2} \right)
\]

\[
\leq \lambda_{\gamma_k} h_k \sqrt{h_k d} \cdot \left( \sqrt{h_k \lambda_{\gamma_k}} + \sqrt{2} \right),
\]

where (i) uses (I83) and (ii) uses (I73). This completes the proof of Lemma 1.

References

Abdulle, A., Vilmart, G., Zygalakis, K.C.: High order numerical approximation of the invariant measure of ergodic SDEs. SIAM J. Numer. Anal. 52(4), 1600–1622 (2014)

Beck, A.: First-Order Methods in Optimization (2017)

Bertsekas, D.: Convex Optimization Theory (2009)

Bou-Rabee, N., Eberle, A.: Markov Chain Monte Carlo Methods (2020)

Brooks, S., Gelman, A., Jones, G., Meng, X.-L.: Handbook of Markov Chain Monte Carlo (2011)

Brosse, N., Durmus, A., Moulines, É., Pereyra, M.: Sampling from a log-concave distribution with compact support with proximal Langevin Monte Carlo. In: Conference on Learning Theory, pp. 319–342. PMLR (2017)

Bubeck, S., Eldan, R., Lehec, J.: Sampling from a log-concave distribution with projected Langevin Monte Carlo. Discrete Comput. Geom. 59(4), 757–783 (2018)

Chambolle, A.: An algorithm for total variation minimization and applications. J. Math. Imaging Vis. 20(1–2), 89–97 (2004)

Dalalyan, A.S.: Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In: Conference on Learning Theory, pp. 678–689. PMLR (2017)

Dalalyan, A.S.: Theoretical guarantees for approximate sampling from smooth and log-concave densities. J. R. Stat. Soc. Ser. B Methodol. 79(3), 651–676 (2017)

Dalalyan, A.S., Karagulyan, A.: User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. Stoch. Process. Their Appl. 129(12), 5278–5311 (2019)

Durmus, A., Moulines, E.: Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. Ann. Appl. Probab. 27(3), 1551–1587 (2017)

Durmus, A., Moulines, E.: High-dimensional Bayesian inference via the unadjusted Langevin algorithm. Bernoulli 25(4A), 2854–2882 (2019)

Durmus, A., Moulines, É., Pereyra, M.: Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau. SIAM J. Imaging Sci 11(1), 473–506 (2018)

Durmus, A., Majewski, S., Miasojedow, B.: Analysis of Langevin Monte Carlo via convex optimization. J. Mach. Learn. Res. 20(73), 1–46 (2019)
Eberle, A., Majka, M.B.: Quantitative contraction rates for Markov chains on general state spaces. Electron. J. Probab. 24 (2019)
Kampf, J.: On weighted parallel volumes. Beiträge Algebra Geom. 50(2), 495–519 (2009)
Klain, D.A., Rota, G.-C.: Introduction to Geometric Probability (1997)
Luu, T.D., Fadili, J., Chesneau, C.: Sampling from non-smooth distributions through Langevin diffusion. Methodol. Comput. Appl. Probab. 23(4), 1173–1201 (2021)
Majka, M.B., Mijatovic, A., Szpruch, L.: Nonasymptotic bounds for sampling algorithms without log-concavity. Ann. Appl. Probab. 30(4), 1534–1581 (2020)
Mattingly, J.C., Stuart, A.M., Tretyakov, M.V.: Convergence of numerical time-averaging and stationary measures via Poisson equations. SIAM J. Numer. Anal. 48(2), 552–577 (2010)
Milstein, G.N., Tretyakov, M.V.: Computing ergodic limits for Langevin equations. Physica D 229(1), 81–95 (2007)
Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Program. 103(1), 127–152 (2005)
Pakman, A., Paninski, L.: Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. J. Comput. Graph. Stat. 23(2), 518–542 (2014)
Pereyra, M.: Proximal Markov chain Monte Carlo algorithms. Stat. Comput. 4(26), 745–760 (2016)
Roberts, G.O., Tweedie, R.L.: Exponential convergence of Langevin distributions and their discrete approximations. Bernoulli 2(4), 341–363 (1996)
Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis (2009)
Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D 60(1–4), 259–268 (1992)
Shepp, L.A., Logan, B.F.: The Fourier reconstruction of a head section. IEEE Trans. Nucl. Sci. 21(3), 21–43 (1974). https://doi.org/10.1109/TNS.1974.6499235
Stella, L., Themelis, A., Patrinos, P.: Forward–backward quasi-newton methods for nonsmooth optimization problems. Comput. Optim. Appl. 67(3), 443–487 (2017)
Villani, C.: Optimal Transport: Old and New (2009)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.