Computationally Efficient Posterior Inference with Langevin Monte Carlo and Early Stopping

Dushyant Sahoo[1,*] Alnur Ali[2,*] Edgar Dobriban[3]

[1]Department of Electrical and Systems Engineering, University of Pennsylvania
[2]Department of Electrical Engineering, and Department of Statistics, Stanford University
[3]Departments of Statistics and Data Science, and of Computer and Information Science, University of Pennsylvania

Abstract

Langevin Monte Carlo is a popular algorithm for posterior inference, and has received growing interest recently because of its simplicity and scalability. A fundamental result, due to Jordan, Kinderlehrer and Otto, reveals that the standard Langevin Monte Carlo method may be seen as a gradient method for maximization, showing a link between optimization and sampling. This insight has been fruitful, and has enabled the application of ideas from optimization to sampling (and vice versa). In this paper, we look at transferring recent results from implicit regularization to sampling. In particular, we prove a number of results showing that an early-stopped variant of the Langevin iteration on the likelihood, can converge to a target posterior faster than the usual Langevin iteration on the posterior. A key feature of our analysis is that we adopt a continuous-time (i.e., stochastic differential equation) perspective, simplifying many of the arguments. We describe some applications of these results, e.g., to Bayesian quadrature, and give extensive numerical evidence supporting our general theory.

Contents

*These authors contributed equally.

1 Introduction

Consider a posterior distribution over some parameters of interest $\beta \in \mathbb{R}^p$ given data $y_1, \ldots, y_n \in \mathbb{R}$, having density $f_{\beta|y_1,\ldots,y_n}$ of the form

$$f_{\beta|y_1,\ldots,y_n}(\beta; y_1, \ldots, y_n) \propto \exp\big(- F_{\beta|y_1,\ldots,y_n}(\beta; y_1, \ldots, y_n)\big),$$

where $F_{\beta|y_1,\ldots,y_n} : \mathbb{R}^p \mapsto \mathbb{R}$. In this paper, we consider the problem of efficiently drawing samples from such a distribution, when $F_{\beta|y_1,\ldots,y_n}$ possesses at least some curvature, but can be quite complicated in general.

Rapidly generating samples from a complex probability model is of course a fundamental problem in both computational statistics and machine learning, and has a number of important applications, with work on the topic going back decades (Ermak, 1975; Cesa-Bianchi and Lugosi, 2006; Rademacher and Vempala, 2008; Braun and McAuliffe, 2010). Broadly speaking, there are two high-level approaches to sampling: Markov chain Monte Carlo-type approaches (Brooks et al., 2011; Robert and Casella, 2013), and variational Bayes-type approaches (Jordan et al., 1999; Blei et al., 2017; Zhang et al., 2018). Variational Bayes methods work by finding a tractable lower bound on the marginal probability of the data that minimizes a measure of divergence from the target posterior, whereas Markov chain Monte Carlo methods work instead by constructing a Markov chain whose stationary distribution is the posterior of interest. Each of these two approaches has its own respective strengths and weaknesses, with the right tool for the job usually depending on the characteristics of the specific problem at hand.

In the current paper, we focus on Markov chain Monte Carlo methods. A key challenge with these is scalability, i.e., as the size of a data set grows, the convergence of Monte Carlo-style methods to the target posterior is notoriously slow. Therefore, in recent years, there has been a considerable push towards developing scalable methods of this type. The class of Hamiltonian Monte Carlo methods (Neal, 2011) — and, in particular, the Langevin Monte Carlo algorithm (Ermak, 1975; Rossky et al., 1978; Parisi, 1981; Grenander, 1983; Neal, 1993; Grenander and Miller, 1994; Roberts and Tweedie, 1996) — have become popular over the last few years because of their simplicity, and good statistical performance across a variety of problems (Mazumdar et al., 2020; Welling and Teh, 2011; Corbineau et al., 2019; De Bortoli et al., 2021).

The standard (i.e., unadjusted, or overdamped) Langevin Monte Carlo algorithm works as follows. Collecting the data $y_1, \ldots, y_n$ into the vector $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, we define the sequence $(\beta^{(k)})_{k \geq 0}$ by initializing $\beta^{(0)} \in \mathbb{R}^p$ sampled from some specified distribution, and by iterating

$$\beta^{(k)} = \beta^{(k-1)} + \tau \cdot \nabla \log f_{\beta|y}(\beta^{(k-1)}; y) + \sqrt{2\tau} \cdot z^{(k)}, \quad \text{for } k = 1, 2, 3, \ldots, \tag{1}$$

where $\tau > 0$ is a fixed step size, and $z^{(k)} \overset{\text{iid}}{\sim} \text{Normal}(0, I_p)$, for all $k \geq 1$. Here and through the paper, $\nabla \log f_{\beta|y}(\cdot; y)$ refers to the gradient of the function $\tilde{\beta} \mapsto \nabla \log f_{\beta|y}(\tilde{\beta}; y)$. At first blush, the iteration (1) above is reminscent of the usual gradient method for maximizing a function $g : \mathbb{R}^p \mapsto \mathbb{R}$, i.e.,

$$\beta^{(k)} = \beta^{(k-1)} + \tau \cdot \nabla g(\beta^{(k-1)}), \quad \text{for } k = 1, 2, 3, \ldots. \tag{2}$$

The connection can, in fact, be made precise; an elegant body of work (see, e.g., Jordan et al. (1998); Dalalyan (2017b,a); Ma et al. (2019b); Wibisono (2018); Ma et al. (2019a); Cheng et al. (2019, 2020); Mou et al. (2021)) elucidates a number of links between variants of the Langevin algorithm and gradient methods.

1.1 Implicit regularization, and this paper

The optimization-based perspective on the standard Langevin algorithm, i.e., viewing (1) and (2) as related in some sense, is fruitful, because it allows us to port ideas from optimization to sampling and

vice versa, e.g., giving rise to new algorithms for sampling (Martin et al., 2012; Simsekli et al., 2016; Brosse et al., 2017; Hsieh et al., 2018; Cheng et al., 2018b; Ma et al., 2019a; Chewi et al., 2020).

To the point, a recent strand of work in optimization has sought to understand the structural properties and "implicit regularization" of the solutions to learning and statistical estimation problems recovered by specific optimization algorithms; see, e.g., Nacson et al. (2019); Gunasekar et al. (2018); Soudry et al. (2018); Suggala et al. (2018); Ali et al. (2019); Poggio et al. (2019); Ji and Telgarsky (2019). A takeaway, translated into the language of the current paper, is as follows (Suggala et al., 2018; Ali et al., 2019, 2020). The maximum a posteriori estimate associated with the posterior

$$f_{\beta|y}(\beta; y) \propto f_{y|\beta}(\beta; y) \cdot f_\beta(\beta),$$

where the prior density $f_\beta(\beta)$ is that of a normal distribution $\text{Normal}(0, I_p/\lambda)$, for some fixed $\lambda > 0$, is equivalent in a certain sense—or rather, close—to the k^*th iterate generated by the gradient ascent iteration (2), with $k^* = \lceil 1/(\tau\lambda) \rceil$, when run on the log likelihood $-F_{y|\beta}$ (corresponding to $g = -F_{y|\beta}$, in (2)).

It is worth re-emphasizing what was just said: ridge-penalized M-estimation is equivalent, in a certain sense, to gradient descent on the negative log likelihood. The connection is an instance of the well-known observation (Strand, 1974; Morgan and Bourlard, 1989; Friedman and Popescu, 2004; Ramsay, 2005; Yao et al., 2007; Raskutti et al., 2014; Wei et al., 2017) that early-stopped gradient descent is equivalent to implicit ℓ_2 regularization. From a practical standpoint, the link is useful because it suggests that we can use the computationally relatively cheap gradient descent iteration, instead of solving the more expensive M-estimation problem.

Naturally, putting the pieces together, we might now ask: can early-stopped Langevin Monte Carlo reach a target posterior faster than the usual Langevin Monte Carlo iteration? The current paper centers around answering this question. In this paper, we show in a precise sense that:

"early-stopped Langevin Monte Carlo on the likelihood $f_{y|\beta}$ is equivalent, in a certain sense, to posterior inference on $f_{\beta|y}$ with a $\text{Normal}(0, I_p/\lambda)$ prior."

We require a bit more notation to make the claim precise; the next section gives a few more details, followed by a summary of our results.

2 A look at the results

A concrete example may help explain the ideas behind the paper. We start by assuming the usual setup in Bayesian linear regression (e.g., Gelman et al. (1995); Bishop (2006); Ghosh et al. (2006)), i.e.,

$$\beta(\lambda) \sim \text{Normal}(0,\, I_p/\lambda) \quad \text{and} \quad y \mid \beta(\lambda) \sim \text{Normal}(X\beta(\lambda),\, \sigma^2 \cdot I_p), \tag{3}$$

where $X \in \mathbb{R}^{n \times p}$ and $\sigma > 0$. In words, we assume both the prior $f_{\beta(\lambda)}$ and the likelihood $f_{y|\beta(\lambda)}$ are Gaussian distributions. In particular, the prior has mean zero and covariance I_p/λ, for some fixed $\lambda > 0$, whereas the likelihood has mean $X\beta$, for a fixed data matrix X and isotropic covariance $\sigma^2 \cdot I_p$. Here and in what follows, we denote the parameters by $\beta(\lambda)$, in order to emphasize their dependence on the prior precision strength λ. The model (3) also implies that the posterior $f_{\beta(\lambda)|y}$ is a Gaussian distribution with a certain mean and covariance structure, i.e.,

$$\beta(\lambda) \mid y \sim \text{Normal}\Big((X^\top X + \lambda \cdot I_p)^{-1} X^\top y,\, (X^\top X + \lambda \cdot I_p)^{-1}\Big). \tag{4}$$

Now the standard Langevin iteration, applied to (3) and starting with some $\beta_\lambda^{(0)}$, is just

$$\beta_\lambda^{(k)} = \beta_\lambda^{(k-1)} + \tau_\lambda \cdot \nabla \log f_{\beta(\lambda)|y}(\beta_\lambda^{(k-1)}; y) + \sqrt{2\tau_\lambda} \cdot z_\lambda^{(k)}, \tag{5}$$

where $\tau_\lambda > 0$ is a step size, and $z_\lambda^{(k)} \sim \text{Normal}(0, I_p)$, for $k = 1, 2, 3, \ldots$. On the other hand, the early-stopped Langevin iteration on just the likelihood $f_{y|\beta(\lambda)}$ in (3), starting with some $\beta^{(0)}$, is

$$\beta^{(k)} = \beta^{(k-1)} + \tau \cdot \nabla \log f_{y|\beta(\lambda)}(\beta^{(k-1)}; y) + \sqrt{2\tau} \cdot z^{(k)}, \tag{6}$$

where $\tau > 0$ and $z^{(k)} \overset{\text{iid}}{\sim} \text{Normal}(0, I_p)$, for all k. Notice the presence of the likelihood $f_{y|\beta(\lambda)}$ in (6), compared with the posterior $f_{\beta(\lambda)|y}$ in (5). Of course, for the Gaussian data model (3), we have that the posterior is of the form

$$f_{\beta(\lambda)|y}(\beta; y) \propto \exp(-\|y - X\beta\|_2^2/2),$$

so that

$$\nabla \log f_{y|\beta(\lambda)}(\beta; y) = X^\top [y - X\beta]. \tag{7}$$

In this paper, we (roughly) claim that

$$W_2^2\big(\text{Law}(\beta^{(k^*)}),\, \text{Law}(\beta(\lambda) \mid y)\big) \approx 0, \tag{8}$$

in a suitable sense, when $k^* = \lceil 1/(\tau\lambda) \rceil$. Here,

$$W_2^2(\text{Law}_1, \text{Law}_2) = \inf_{P \in \mathcal{P}(\text{Law}_1, \text{Law}_2)} \mathbb{E}_{W,Z \sim P} \|W - Z\|_2^2$$

in (8) denotes the squared 2-Wasserstein distance between the laws—i.e., probability distributions—Law_1, Law_2, and $\mathcal{P}(\text{Law}_1, \text{Law}_2)$ denotes all couplings of those laws, namely joint probability distributions of (W, Z) having Law_1, Law_2 as the marginals of W, Z, respectively (e.g., Villani, 2003). We frequently write $W_2^2(W,\, Z)$ as a short-hand for the same quantity. The Wasserstein distance is an intuitive and convenient notion of distance on the space of probability measures (e.g., absolute continuity restrictions), and bears connections to other well-known probability metrics. However, the proposal above, i.e., doing early stopping on the process (6) instead of running (5) to convergence, may seem somewhat strange at first, because of issues related to burn-in. The key is to recognize that we must perform early stopping on a *different* Markov chain, i.e., one associated with the likelihood (not the posterior, as we usually would).

Moreover, because the laws are close, the expectations must also be close, and therefore we also have, for any Lipschitz continuous test function $g : \mathbb{R}^p \mapsto \mathbb{R}$, that

$$\frac{1}{k^*} \sum_{j=1}^{k^*} g(\beta^{(j)}) \approx \frac{1}{\ell} \sum_{j=1}^{\ell} g(\beta_\lambda^{(j)}) \approx \mathbb{E}_{\beta(\lambda)|y}\big[g(\beta(\lambda))\big],$$

in a suitable sense, where $\ell \gg k^*$ is some (large) number of iterations.

A short numerical example. Returning to the main thread, we may check the claim (8) above by computing the relative sampling efficiency,

$$\text{RSE}(\beta^{(k)},\, \beta_\lambda^{(k)}) = \frac{W_2^2\big(\text{Law}(\beta_\lambda^{(k)}),\, \text{Law}(\beta(\lambda) \mid y)\big)}{W_2^2\big(\text{Law}(\beta^{(k)}),\, \text{Law}(\beta(\lambda) \mid y)\big)}, \tag{9}$$

so that the relative sampling efficiency is large when the denominator is smaller than the numerator, and hence early stopping on $(\beta^{(k)})_{k \geq 1}$ helps. This is what we would expect around $k^* = \lceil 1/(\tau\lambda) \rceil$ iterations, in line with (8). Figure 1 shows the results of a small simulation study, where we generated data according to (3) for three different values of $\lambda \in \{0.1, 0.3, 0.5\}$ (see Section 7 for details), ran the discrete time iterations (5), (6), and then computed (9). From the figure, we can see that the relative sampling efficiency becomes large near k^* iterations, as expected.

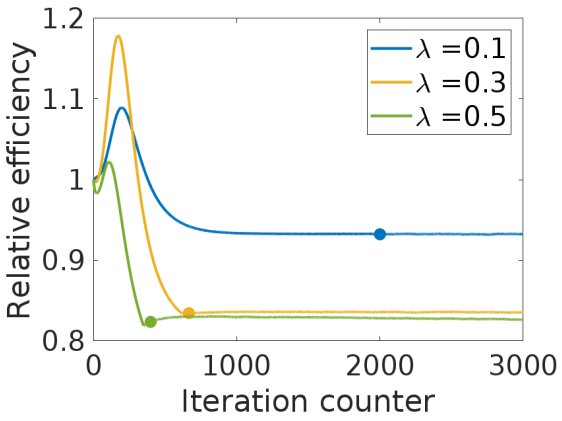

Figure 1: *The relative sampling efficiency, defined in* (9)*, of the early-stopped Langevin Monte Carlo iteration on the likelihood* (6) *over the standard Langevin Monte Carlo iteration on the posterior* (5)*, for Bayesian linear regression* (4)*, with various values of the prior precision strength* $\lambda \in \{0.1, 0.3, 0.5\}$*. The relative sampling efficiency is large when the number of iterations* $k \approx 1/(\tau\lambda)$*, as indicated by the dots, and small otherwise, suggesting that early-stopped Langevin Monte Carlo on the likelihood is in fact performing posterior inference, as our theory predicts. See Sections 4 and 7, for further details.*

2.1 Overview of contributions

Below is a summary of our contributions in this paper.

- In Section 4, we give a tight bound on the squared 2-Wasserstein distance between the law of the early-stopped Langevin iteration on the likelihood at $k^* = \lceil 1/(\tau\lambda) \rceil$ iterations, i.e., $\text{Law}(\beta_{k^*})$, and the law of the true posterior over $\beta(\lambda)$, under the Gaussian data model (3).

- In Section 5, we move beyond the fundamental Gaussian model, and give a few more general bounds on the Wasserstein distance between $\beta^{(k)}$ and $\beta(\lambda)$, where we require that the prior still be Gaussian, but the likelihood need not be. First, we give a bound for the situation when the likelihood is strongly log-concave (in $\beta(\lambda)$). Then, we present a bound covering the more general case when the likelihood is only strongly log-concave outside of a neighborhood centered around the prior mean.

- We also provide a related bound on the Monte Carlo integration mean squared error in Section 6, i.e.,

$$\mathbb{E}_{\beta(\lambda)|y}\left[\left(\frac{1}{k^*}\sum_{j=1}^{k^*} g(\beta^{(j)}) - \mathbb{E}_{\beta(\lambda)|y}\big[g(\beta(\lambda))\big]\right)^2\right],$$

  where $g : \mathbb{R}^p \mapsto \mathbb{R}$ is a smooth function, arising from using early-stopped Langevin dynamics for numerical integration. Numerical integration is, of course, an important and challenging problem intrinsically tied to posterior inference; our bound (as well as the associated numerical experiments) illustrate that the early-stopped iteration can help here, too.

- Finally, in Section 7, we present extensive numerical simulations — covering Bayesian generalized linear models, quadrature, and non-strongly-log-concave likelihoods — giving backing to our general theory.

2.2 Outline

Here is a brief outline for the rest of this paper. We review related work in the next section. Then, we study the Bayesian linear regression setting discussed above in more detail, and give a few different

quantitative results, in Section 4. In Section 5, we relax the assumption that the likelihood must be Gaussian, and allow it to be strongly log-concave outside of a ball. Finally, in Section 7, we present our numerical experiments, before concluding with a short discussion in Section 8.

3 Related work

There is a lot of work related to the ideas in this paper. Sampling is, of course, an old topic, with a rich history in both applied mathematics and statistics. The introduction of the Langevin Monte Carlo method itself can be traced as far back as Ermak (1975); Rossky et al. (1978); Parisi (1981); Grenander (1983); Neal (1993); Grenander and Miller (1994); Roberts and Tweedie (1996) (at least). Some early works studying the method, focusing on its convergence properties, include Jordan et al. (1998); Talay and Tubaro (1990); Gelfand and Mitter (1991); Lamberton and Pages (2002). More recently, the Langevin Monte Carlo method has received a resurgence of interest, driven at least partly by a growing preference for scalable algorithms; these recent works have mainly looked at providing non-asymptotic convergence analyses under various notions of divergence, and under various assumptions (e.g., smoothness and convexity) tailored for modern applications. Some key references here include (but this is by no means an exhaustive list): Dalalyan and Tsybakov (2009); Dalalyan (2017a,b); Zou et al. (2020); Dalalyan and Riou-Durand (2020); Cheng et al. (2018a); Cheng and Bartlett (2018); Cheng et al. (2018b); Hodgkinson et al. (2021); Ma et al. (2019b); Dalalyan and Karagulyan (2019); Dalalyan et al. (2019); Cheng et al. (2019); Ma et al. (2019a); Erdogdu and Hosseinzadeh (2021); Nguyen et al. (2021). On a related note, recent work (Welling and Teh, 2011; Sato and Nakagawa, 2014; Teh et al., 2016; Raginsky et al., 2017; Brosse et al., 2018) introduced and studied stochastic gradient Langevin dynamics, a scalable and popular variant of the standard Langevin Monte Carlo method.

Analyses of the basic Langevin iteration performed in continuous-time — also a key feature of our approach in this paper, as we detail in the next section — appeared immediately; see, e.g., Villani (2008); Ambrosio et al. (2008); Gozlan and Léonard (2010). Continuous-time analyses have seen wide applicability in various corners of convex optimization (Raginsky and Bouvrie, 2012; Xu et al., 2018; Orvieto and Lucchi, 2019), and more recently, statistical theory (Mandt et al., 2015; Wang and Wu, 2020; Ali et al., 2019, 2020).

The optimization-based view of sampling algorithms can be traced back to the seminal work of Jordan et al. (1998), with plenty of recent follow-up, e.g., Martin et al. (2012); Simsekli et al. (2016); Brosse et al. (2017); Hsieh et al. (2018); Cheng et al. (2018b); Ma et al. (2019a); Chewi et al. (2020).

Finally, implicit regularization has, of course, seen intense activity over the last few years; see, e.g., Yao et al. (2007); Neu and Rosasco (2018); Du et al. (2018); Gunasekar et al. (2018); Nacson et al. (2019); Wu et al. (2020); Vaskevicius et al. (2019). Most of these papers study the statistical properties of the convergence points of various iterative optimization algorithms, though there are a few exceptions that study pathwise behavior (Suggala et al., 2018; Ali et al., 2019, 2020) — making them very relevant to the analyses we carry out in the current paper.

4 Bayesian linear regression

We begin by analyzing the Gaussian setup, first presented in (3), in more detail, and make precise the qualitative observations we made in Section 2. The key to the analysis is adopting a continuous-time perspective. It is well-known that the Langevin iterations given in (5), (6) are the Euler discretizations of the continuous-time stochastic processes $(\tilde{\beta}_t)_{t\in[0,\infty)}$, with initialization $\tilde{\beta}_0 = \beta_\lambda^{(0)}$ and

$$d\tilde{\beta}_t = \nabla \log f_{\beta(\lambda)|y}(\tilde{\beta}_t; y)dt + \sqrt{2} \cdot d\tilde{W}_t \tag{10}$$

and $(\beta_t)_{t\in[0,\infty)}$, with initialization $\beta_0 = \beta^{(0)}$ and

$$d\beta_t = \nabla \log f_{y|\beta(\lambda)}(\beta_t; y)dt + \sqrt{2} \cdot dW_t, \tag{11}$$

respectively. Here, $(\tilde{W}_t)_{t\in[0,\infty)}$ and $(W_t)_{t\in[0,\infty)}$ denote two independent instances of standard p-dimensional Brownian motion. We refer to the processes (10), (11) as standard and early-stopped Langevin dynamics, respectively. We can always apply our results in continuous-time to the original discrete time Langevin iterations (5), (6), as due to the triangle inequality

$$W_2\big(\mathrm{Law}(\beta^{(k)}),\, \mathrm{Law}(\beta(\lambda) \mid y)\big) \leq W_2\big(\mathrm{Law}(\beta^{(k)}),\, \mathrm{Law}(\beta_t)\big)\big) + W_2\big(\mathrm{Law}(\beta_t),\, \mathrm{Law}(\beta(\lambda) \mid y)\big), \tag{12}$$

and the first term on the right-hand side simply accounts for discretization error; cf. Corollary 4, appearing below.

In the Gaussian setting, i.e., from (7), the dynamics in (11) simply become

$$d\beta_t = X^\top[y - X\beta_t]dt + \sqrt{2} \cdot dW_t. \tag{13}$$

Standard results (see, e.g., Kloeden and Platen (2013), or Lemma 2 in Ali et al. (2020)) then imply that the stochastic differential equation (13) has a unique strong solution. The Gaussian setting, in particular, is convenient because the dynamics in (13) belong to the class of linear stochastic differential equations, i.e., differential equations that are linear in both the process β_t itself, as well as the increments of the Brownian motion dW_t. The Ornstein-Uhlenbeck process is a well-known example of this kind of stochastic process. Good general references on the topic are Øksendal (2003); Shreve (2004).

We can always obtain an explicit expression for the solution β_t to a linear stochastic differential equation. For (13), we have, for any fixed $T > 0$ and for all $t \in [0, T]$,

$$\beta_t = \gamma_t + \int_0^t \exp\big((s-t)X^\top X\big)\sqrt{2}dW_s, \tag{14}$$

where

$$\gamma_t = (X^\top X)^+\big(I - \exp(-tX^\top X)\big)X^\top y,$$

A^+ denotes the Moore-Penrose pseudo-inverse of A, and $\exp(A)$ denotes the matrix exponential of A; additionally, see Lemma 2 in the appendix. To be clear, the randomness in the process β_t arises from the random parameters $\beta(\lambda)$, as well as the randomness inherent to the Langevin Monte Carlo algorithm itself. Combining (14) with a convenient representation of the Wasserstein distance between two Gaussians (see, e.g., Proposition 7 in Givens and Shortt (1984), or Takatsu (2011)) leads to the following result, tightly coupling the laws of the early-stopped Langevin dynamics β_t and the underlying parameters $\beta(\lambda) \mid y$, in expectation over draws of y.

When stating our results in this paper, we always write $s_i \geq 0$, $i = 1, \ldots, p$, for the eigenvalues of the Gram matrix $X^\top X$. For two real-valued variables a, b, we write $a \wedge b = \min\{a, b\}$. For a positive integer i, we write $[i] = \{1, 2, \ldots, i\}$. We also write $L = \max_{i\in[p]} s_i$, and $m = \min_{i\in[p]} s_i$, for the largest and smallest eigenvalues of the Gram matrix, respectively. Throughout the current section, we assume that $m > 0$, though this assumption could be relaxed here, at the expense of more complicated proofs.

Theorem 1 (Bound on expected W_2^2)**.** *Let $A = 0.2379$ and $B = 1.026$. Under the model* (3)*, the law of the early-stopped Langevin dynamics* (13)*, at the stopping time $t^* = 1/\lambda$ and under the initialization $\beta_0 = 0$, satisfies*

$$\mathbb{E}_{\beta(\lambda),y}\big[W_2^2(\beta_{t^*},\, \beta(\lambda) \mid y)\big] \leq \sum_{i=1}^{p} \frac{(As_i) \wedge (B\lambda)}{s_i(s_i + \lambda)}.$$

We make a few remarks on the theorem.

- The bound in Theorem 1 says that, in general, we should expect the Wasserstein distance between β_t and $\beta(\lambda) \mid y$ to be small. At the extreme points, e.g., when the prior precision strength $\lambda \to \infty$, we can see that the bound goes to zero. Intuitively, this makes sense, as in this case the posterior from (4) is concentrated at the prior mean (i.e., zero), and the initial point $\beta_t = 0$ is optimal. Similarly, when $\lambda \to 0$, the prior is flat, i.e., uninformative, so that β_t with $t = 1/\lambda \to \infty$ is optimal as it converges in distribution to the least squares estimator, (4) with $\lambda = 0$.

- We believe the emergence of the (small) absolute constants $A = 0.2379$ and $B = 1.026$ is rather remarkable. These constants come from numerically maximizing a certain function that arises in the proof. Additionally, another interesting feature of Theorem 1 is that it reduces a complex object of interest, i.e., the Wasserstein distance with respect to a continuous-time stochastic process, to a simple functional of the eigenvalues of the Gram matrix, with explicit numerical constants.

- At a high-level, the proof strategy is to first invoke the explicit representation of the solution β_t, as in (14), to the linear stochastic differential equation (13), and its Wasserstein distance to $\beta(\lambda) \mid y$. Then, we carefully control the functionals of the eigenvalues s_i, $i = 1, \ldots, p$, that follow from these expressions, in order to get a tight bound.

- In Section 7, we numerically investigate the tightness of the bound from Theorem 1.

It is also possible to give a similar result bounding the distance between the laws of β_t and $\beta(\lambda) \mid y$, except holding with high probability; the following corollary gives the details.

Corollary 1 (Bound on W_2^2, high probability version)**.** *Assume the same conditions as in Theorem 1. Let*

$$A_{t,\lambda} = Q_{t,\lambda} \left[\begin{array}{cc} \lambda^{-1/2} X & I_n \end{array} \right],$$
$$Q_{t,\lambda} = \left[(I_p - \exp(-tX^\top X))(X^\top X)^+ - (X^\top X + \lambda I_p)^{-1} \right] X^\top.$$

Then, the law of the early-stopped Langevin dynamics (13)*, at the stopping time $t^* = 1/\lambda$ and under the initialization $\beta_0 = 0$, satisfies*

$$W_2^2(\beta_{t^*},\, \beta(\lambda) \mid y) \leq \sum_{i=1}^{p} \frac{(As_i) \wedge (B\lambda)}{s_i(s_i + \lambda)} + w,$$

with probability (taken over the randomness in y) at least

$$1 - 2\exp\left[-c \min\left(\frac{w^2}{\|A_{t,\lambda}\|_4^4}, \frac{w}{\|A_{t,\lambda}\|_\infty^2} \right) \right],$$

for a universal constant $c > 0$ and all $w > 0$.

In light of Theorem 1 and Corollary 1, a natural question is under what conditions the early-stopped Langevin iteration on the likelihood can reach the target posterior faster than the standard Langevin iteration on the posterior. To make the question precise, we can equivalently ask for sufficient conditions under which the early-stopped dynamics β_{t^*}, i.e., the dynamics at time $t^* = 1/\lambda$, attains smaller Wasserstein distance to $\beta(\lambda) \mid y$ than the standard dynamics $\tilde{\beta}_{t^*}$ does, at the same time t^*. Because the computational cost of the two discrete time iterations is comparable (both cost $O(np)$, when $n \gg p$), this would imply that the early-stopped iteration is more efficient than the standard iteration, given an a priori fixed computational budget. In the following lemma, we spell out sufficient conditions for the early-stopped dynamics to be more efficient than the standard dynamics.

Lemma 1 (Sufficient condition for faster convergence)**.** *Assume the same conditions as in Theorem 1. Write $x = L/\lambda$, and fix the prior precision λ so that it satisfies*

$$\lambda \leq \frac{\exp(-2[x+1])}{A}.$$

Initialize the standard continuous-time Langevin dynamics (10) *with $\tilde{\beta}_0 = 0$. Then, for $t^* = 1/\lambda$,*

$$\mathbb{E}_{\beta(\lambda),y}\big[W_2^2(\beta_{t^*},\, \beta(\lambda) \mid y)\big] \leq \mathbb{E}_{\beta(\lambda),y}\big[W_2^2(\tilde{\beta}_{t^*},\, \beta(\lambda) \mid y)\big].$$

That is, the early-stopped Langevin dynamics reaches the target posterior faster than the standard Langevin dynamics does. The range of λ prescribed by the lemma calls for the prior to be relatively concentrated, which is an important regime in practice. However, in our experiments to come in Section 7, we observe faster convergence for a wide range of λ (i.e., for both small and large values of λ).

4.1 The implicit prior, and empirical Bayes

Thus far, we have argued that early-stopped Langevin dynamics, using only the likelihood $f_{y|\beta(\lambda)}$, can be more efficient than standard Langevin dynamics, which uses the likelihood as well as the prior $f_{\beta(\lambda)}$. Therefore, our claim may appear a little surprising, as we might worry that early stopping is, in a sense, "discarding" information. We think there are two possible resolutions to this tension. First, the early-stopped Langevin iteration (6) does in fact use information about the prior through the choice of the stopping time $t^* = 1/\lambda$.

On the other hand, it turns out that early stopping can be seen as operating with a certain prior, *implicitly*; we give some details now, in order to bring out the connection. From the explicit representation of the solution β_t to the early-stopped process, given in (14), we can see that β_t may be decomposed into two parts: a deterministic part γ_t, and a white noise part. Moreover, Lemma 4 in Ali et al. (2020) shows that the deterministic part can be seen as the solution to a quadratically regularized least squares problem, i.e., we have that

$$\gamma_t = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2}\|y - X\beta\|_2^2 + \beta^\top Q_t \beta \right\}, \quad \text{where} \quad Q_t = VS\big(\exp(tS) - I\big)^{-1}V^\top,$$

the columns of $V \in \mathbb{R}^{n\times p}$ contain the eigenvectors of $X^\top X$, and the diagonal entries of $S \in \mathbb{R}^{p\times p}$ contain the eigenvalues with zeros everywhere else. But this is equivalent to finding the maximum a posteriori estimate under the prior

$$\beta \sim \text{Normal}(0, \sigma^2 Q_t^{-1}),$$

instead of the prior in (3). The (implicit) prior on β here has an intriguing interpretation: it depends not only on t, but also on the data matrix X (through its empirical covariance matrix), reminiscent of empirical Bayes-type methods (Ghosh et al., 2006; Efron, 2012).

5 Sampling from a generic posterior

5.1 Strongly log-concave likelihood

In this section, we move beyond the canonical Gaussian setup, and assume a bit more for the posterior of interest. Namely, we consider a data-generating process where the prior is still normal, but the likelihood need not be, as in

$$\beta(\lambda) \sim \text{Normal}(0,\, I_p/\lambda) \quad \text{and} \quad y \mid \beta(\lambda) \sim f_{y|\beta(\lambda)}(\beta(\lambda); y), \tag{15}$$

where in particular the log likelihood $-F_{y|\beta(\lambda)} = \log f_{y|\beta(\lambda)}$ must satisfy the following two assumptions. Here and throughout the paper, we generally omit any normalizing factors when writing down densities, in order to keep the notation light.

Assumption A1 (Lipschitz smoothness)**.** The gradient of the log likelihood $\nabla \log f_{y|\beta(\lambda)}$ is L-Lipschitz continuous uniformly in $y \in \mathbb{R}^n$, i.e., there exists some $L > 0$, such that for all $\theta_1, \theta_2 \in \mathbb{R}^p$, $y \in \mathbb{R}^n$

$$\|\nabla \log f_{y|\beta(\lambda)}(\theta_1; y) - \nabla \log f_{y|\beta(\lambda)}(\theta_2; y)\|_2 \leq L \cdot \|\theta_1 - \theta_2\|_2.$$

For a positive integer i and two $i \times i$ symmetric matrices M_1, M_2, we write $M_1 \succeq M_2$ when $M_1 - M_2$ is positive semi-definite.

Assumption A2 (Strong log-concavity)**.** The likelihood $f_{y|\beta(\lambda)}(\cdot; y)$ is m-strongly log-concave uniformly in $y \in \mathbb{R}^n$, i.e., there exists some $m > 0$, such that for all $\theta \in \mathbb{R}^p$, $y \in \mathbb{R}^n$

$$\nabla^2 \log f_{y|\beta(\lambda)}(\theta; y) \succeq m \cdot I_p.$$

Assumptions A1 and A2 are both reasonably broad, and also standard in the literature on optimization (Nesterov, 2003) as well as sampling; see, e.g., Cheng and Bartlett (2018); Lee et al. (2019); Karagulyan and Dalalyan (2020). For instance, they clearly capture the Gaussian setting of (13), though the bounds we develop based on these assumptions are looser than the one we presented in, say, Theorem 1. In the special Gaussian setting, they concide with the previous definitions of L, m. Moreover, these conditions imply (see, e.g., Proposition 6.1 in Pavliotis (2014)) that Langevin dynamics (11) converges to an invariant distribution π.

With these assumptions in hand, we can prove the following result, controlling the Kullback-Leibler divergence between the laws of β_t, as in (11), and the posterior $\beta(\lambda) \mid y$, arising from (15). In what follows, we always write μ for the mean of the invariant distribution π of the dynamics (11).

Corollary 2 (Bound on expected KL divergence, strongly log-concave likelihood)**.** *Assume the data model* (15)*, as well as conditions A1 and A2, stated above. Define the constants*

$$\begin{aligned} a_\lambda &= \tfrac{Lp}{2\lambda} + \tfrac{p}{2} \log \tfrac{\lambda}{m+\lambda}, & m'_\lambda &= m + \lambda, \\ b &= \tfrac{2p}{m} + 2\|\mu\|_2^2, & b_\lambda &= \tfrac{2p}{m} \cdot \left(\tfrac{L-\lambda}{\lambda} + \log \lambda/m \right). \end{aligned}$$

Then, the law of early-stopped Langevin dynamics (11)*, at the stopping time $t^* = 1/\lambda$ and under the initialization $\beta_0 \sim \mathrm{Normal}(0, I_p/\lambda)$, satisfies*

$$\begin{aligned} \mathbb{E}_{\beta(\lambda),y}\big[D_{\mathrm{kl}}(\beta_{t^*} \| \beta(\lambda) \mid y)\big] \leq a_\lambda \cdot \exp(-m'_\lambda/\lambda) &+ \frac{b\lambda^2}{2} \cdot \frac{1 - \exp(-m'_\lambda/\lambda)}{m'_\lambda} \\ &+ \frac{b_\lambda \lambda^2}{2} \cdot \frac{\exp\big((m'_\lambda - 2m)/\lambda\big) - 1}{m'_\lambda - 2m} \cdot \exp(-m'_\lambda/\lambda). \end{aligned}$$

We prove a more general result in the appendix (see Theorem 4) that implies both the above result and a more general one to come (see Corollary 3 below) as special cases, but we present the results separately for readability. A few remarks on Corollary 2 are in order.

- The bound in Corollary 2 is more general than the bound in Theorem 1; of course, it is also somewhat looser. We investigate this trade-off through several numerical experiments in Section 7.
- The Gaussian initialization in the above bound, which is different from that in Theorem 1, is due to technical considerations arising in the proof; see the proof of Theorem 1 in the appendix for a discussion.

- The proof strategy is essentially to control the derivative,

$$\frac{dD_{\mathrm{kl}}(\beta_t \| \beta(\lambda) \mid y)}{dt},$$

by using tools from the classical theory of gradient flows; see Ambrosio et al. (2008) for a treatment of this subject. We then apply Gronwall's inequality in order to obtain control on the quantity $D_{\mathrm{kl}}(\beta_t \| \beta(\lambda) \mid y)$ itself. Finally, we simplify the resulting bound, by using the correspondence $t = 1/\lambda$, and working out bounds on a few intermediate quantities, e.g., (i) $\mathbb{E}_{\beta(\lambda),y} \|\beta_t\|_2^2$, (ii) $D_{\mathrm{kl}}(\beta_0 \| \beta(\lambda) \mid y)$, and (iii) the Kullback-Leibler divergence between β_0 and the stationary distribution of the early-stopped process.

- As $\lambda \to 0$, it can be checked that the bound converges to zero, as expected. Similarly, as m grows, the log density has more curvature, and we can see that the bound shrinks in this case, too (as expected). On the other hand, a weakness of this bound is that it does not converge to zero, as $\lambda \to \infty$. Later, we give a more sophisticated bound that is sharper, but is less transparent than the current bound; see Theorem 2.

5.1.1 A transportation cost inequality

Finally, converting a bound on the Kullback-Leibler divergence into the Wasserstein distance follows readily, provided the posterior satisfies a transportation cost (also often referred to as a Talagrand-type) inequality (Talagrand, 1996; Otto and Villani, 2000; Boucheron et al., 2013), described next.

We say that the probability measure P satisfies a transportation cost inequality with constant α, if for all probability measures $Q \ll P$ absolutely continuous with respect to P,

$$W_2^2(Q, P) \leq \frac{2}{\alpha} \cdot D_{\mathrm{kl}}(Q \| P). \tag{16}$$

Talagrand (1996) first proved an inequality of this type for Gaussian random variables; the result was later extended by Otto and Villani (2000) to any probability measure satisfying a log Sobolev inequality, or equivalently being strongly log-concave, which is certainly the case in (3). Since Assumption A2 implies the transportation cost inequality (16) is satisfied with constant m'_λ, we immediately obtain a bound on the Wasserstein distance, i.e., we have

$$\mathbb{E}_{\beta(\lambda),y}\big[W_2^2(\beta_{t^*},\, \beta(\lambda) \mid y)\big] \leq \frac{2}{m'_\lambda} \cdot \mathbb{E}_{\beta(\lambda),y}\big[D_{\mathrm{kl}}(\beta_{t^*} \| \beta(\lambda) \mid y)\big].$$

The inequality (16) and the log-Sobolev inequality turn out to be critical for the remainder of this paper, and we return to them in the next section.

5.2 Beyond a strongly log-concave likelihood

Now we additionally assume that the target posterior satisfies a log Sobolev inequality with constant $\alpha > 0$. We also relax Assumption A2, allowing the density to only be strongly log-concave outside of a neighborhood of the prior mean. These requirements are stated formally below.

Assumption A3 (Strong log-concavity outside of a ball)**.** There exists a p-dimensional Euclidean ball $\mathcal{B}_0^p(R)$ centered around zero with radius $R > 0$, such that Assumption A2 is satisfied when Θ restricted to $\mathrm{dom}(f_{y|\beta(\lambda)}(\cdot; y)) \setminus \mathcal{B}_0^p(R)$.

Assumption A4 (Log Sobolev inequality)**.** The likelihood $f_{y|\beta(\lambda)}(\cdot; y)$ satisfies a log Sobolev inequality, i.e., there exists some constant $\alpha > 0$, such that for all square integrable and Lipschitz smooth functions $g : \mathbb{R}^p \mapsto \mathbb{R}$, taking expectations with respect to the random vector $W \in \mathbb{R}^p$, $W \sim f_{y|\beta(\lambda)}$,

$$\mathbb{E}\big[g^2(W) \log g^2(W)\big] - \mathbb{E}\big[g^2(W) \log \mathbb{E} g^2(W)\big] \leq \mathbb{E}\Big[\frac{2}{\alpha} \cdot \|\nabla g(W)\|_2^2\Big].$$

The largest α satisfying the above inequality is called the log Sobolev constant. This inequality was first established in Gross (1975), for Gaussian random variables. It was later shown by Bakry and Émery (1985) to hold for any random variable following a strongly log-concave distribution. A log Sobolev inequality of the above type essentially relates the entropy of a random variable to its gradient, and therefore bounds the fluctuations of the random variable, making its appearance here relatively natural. The log Sobolev inequality also has deep connections to concentration of measure; via the Efron-Stein inequality, maximal inequalities for stochastic processes, and isoperimetric inequalities (see, e.g., Boucheron et al. (2013), for a review). A convenient feature of the log Sobolev inequality is that it is preserved under both products and Lipschitz transformations. It also always implies the inequality,

$$D_{\mathrm{kl}}(P\|Q) \leq \frac{1}{2\alpha} \cdot \mathbb{E}_P \left\| \nabla \log \frac{dP}{dQ} \right\|_2^2, \tag{17}$$

which we exploit in the proof of our result. The inequality (17) can be seen by substituting $g = \sqrt{dP/dQ}$, and then performing some algebra. We state our KL divergence result under these assumptions next.

Corollary 3 (Bound on expected KL divergence, non-strongly-log-concave posterior)**.** *Assume the data model* (15)*, as well as conditions A1, A3, and A4, given above. Define the constants*

$$a_\lambda = \tfrac{Lp}{2\lambda}, \qquad m'_{R,\lambda} = \lambda,$$
$$b_R = \tfrac{16p}{\alpha} \log 2L/m + \tfrac{512R^2L^3}{\alpha m^2} + 2\|\mu\|_2^2, \qquad b_{R,\lambda} = \tfrac{2p}{\alpha} \cdot \left(\tfrac{L-\lambda}{2\lambda} + \tfrac{1}{2} \log \tfrac{2\lambda}{m} + \tfrac{32R^2L^3}{pm^2} \right).$$

Then, the law of early-stopped Langevin dynamics (11)*, at the stopping time $t^* = 1/\lambda$ and under the initialization $\beta_0 \sim \mathrm{Normal}(0, I_p/\lambda)$, satisfies*

$$\mathbb{E}_{\beta(\lambda),y}\big[D_{\mathrm{kl}}(\beta_{t^*}\|\beta(\lambda) \mid y)\big] \leq a_\lambda \cdot \exp(-m'_{R,\lambda}/\lambda) + \frac{b_R\lambda^2}{2} \cdot \frac{1-\exp(-m'_{R,\lambda}/\lambda)}{m'_{R,\lambda}}$$
$$+ \frac{b_{R,\lambda}\lambda^2}{2} \cdot \frac{\exp\big((m'_{R,\lambda}-2\alpha)/\lambda\big)-1}{m'_{R,\lambda}-2\alpha} \cdot \exp(-m'_{R,\lambda}/\lambda).$$

The difference between the bounds given in Corollaries 3 and 2 lies in the definitions of their respective constants. Crucially, the bound from Corollary 2 does not simply follow from setting $R = 0$ in Corollary 3.

5.2.1 A discrete time result

Now we return to a point raised at the beginning of the paper, and translate Corollary 3 presented above into discrete time.

Corollary 4 (Bound on expected W_2^2, discrete time)**.** *Assume the same conditions as in Corollary 3. Let L_ξ be an arbitrarily small constant, $c = \sqrt{2}$, and additionally define the constants*

$$\eta = min\left\{\tfrac{m}{2}, \tfrac{1}{8R^2}\right\} \exp\left(-\tfrac{7}{12}LR^2\right), \qquad \hat{\epsilon} \leq \left(\tfrac{16L}{\eta}\right) \exp(7LR/12)^{\frac{R}{LR^2/4+1}}.$$

Then, there exists a coupling between the laws of discrete and continuous-time early-stopped Langevin dynamics $\beta^{(k)}$ and β_t, as in (6) *and* (11)*, respectively, with step size satisfying*

$$\tau \leq \min \begin{cases} \frac{\eta^2\hat{\epsilon}^2}{512\sqrt{p}L^2 \exp\left(\frac{7L}{2R^2}\right)} \\ \frac{2\eta\hat{\epsilon}}{\sqrt{p^2/\lambda}L^2 \exp\left(\frac{7L}{4R^2}\right)} \end{cases},$$

and initialized at $\beta^{(0)} = \beta_0$ *with* $\mathbb{E}\left[\|\beta_0\|_2^2\right] \leq \frac{p}{\lambda}$*, such that* $\beta^{(k)}$ *and* β_t *satisfy at* $k^* = \lceil 1/(\tau\lambda) \rceil$ *iterations and time* t^*

$$\mathbb{E}\left[\|\beta^{(k^*)} - \beta_{t^*}\|_2\right] \leq \hat{\epsilon}.$$

The proof of the result follows by controlling the discretization error term in (12). Fortunately, this is possible, by extending a recent result due to Cheng et al. (2020) coupling the laws of the Langevin iteration with the target posterior $\beta(\lambda) \mid y$; see Section 10.1.6 in the appendix for further details.

5.3 A sharper bound

We now present a final bound that sharpens the previous continuous-time results. On the flipside, the bound is less transparent than our earlier bounds. Roughly speaking, we prove the new bound by optimizing the stopping time t^* in order to make a certain functional arising in the proof of the result as small as possible; therefore, the new bound also does not explicitly assume the relation $t^* = 1/\lambda$.

Theorem 2 (Sharper bound on expected KL divergence, non-strongly-log-concave posterior)**.** *Assume the same conditions as in Corollary 3, and recall the definitions of* $a_\lambda, m'_{R,\lambda}, b_R, b_{R,\lambda}$ *from that result. For* $a \in (0,2)$*, define,*

$$\begin{aligned} m''_{R,\lambda,a} &= 2(1-a/2)m'_{R,\lambda}, & v_{R,\lambda,a} &= \tfrac{b_{R,\lambda}\lambda^2}{2a(m''_{R,\lambda,a}-\alpha)}, \\ q_{R,\lambda,a} &= \tfrac{b_R\lambda^2}{2m''_{R,\lambda,a}a}, & r_{R,\lambda,a} &= a_\lambda - q_{R,\lambda,a} - v_{R,\lambda,a}. \end{aligned}$$

Then, the law of early-stopped Langevin dynamics (11)*, at the stopping time*

$$t^* = \begin{cases} \max\left(\frac{1}{\alpha - m''_{R,\lambda,a}} \cdot \log\left(\frac{-\alpha v_{R,\lambda,a}}{r_{R,\lambda,a} m''_{R,\lambda,a}}\right), 0\right), & \text{if } \frac{\alpha v_{R,\lambda,a}}{r_{R,\lambda,a} m''_{R,\lambda,a}} < 0 \\ \infty, & \text{otherwise} \end{cases}$$

and under the initialization $\beta_0 \sim \mathrm{Normal}(0, I_p/\lambda)$*, satisfies*

$$\mathbb{E}_{\beta(\lambda),y}\big[D_{\mathrm{kl}}(\beta_{t^*} \| \beta(\lambda) \mid y)\big] \leq \inf_{a \in (0,2)} G(a; \lambda, L, p, m),$$

where

$$G(a; \lambda, L, p, m) = q_{R,\lambda,a} + r_{R,\lambda,a} \cdot \exp(-m''_{R,\lambda,a} t^*) + v_{R,\lambda,a} \cdot \exp(-\alpha t^*).$$

6 Monte Carlo integration

A related problem, arising in the context of Bayesian inference but also prevalent throughout applied mathematics, is to compute a (possibly very high-dimensional) conditional expectation, with respect to a given prior distribution. Formally, we seek to approximately compute

$$\int_\beta g(\beta) f_{\beta(\lambda)|y}(\beta) d\beta = \mathbb{E}_{\beta(\lambda)|y}\big[g(\beta(\lambda))\big], \tag{18}$$

where $g : \mathbb{R}^p \mapsto \mathbb{R}$ is a given Lipschitz continuous "test" function. In general, computing the expectation in (18) is difficult. The usual approach is to use some variation of the standard Monte Carlo method, which in its most basic form performs the following two steps:

$$\begin{aligned} &\text{draw:} && \beta^{(i_1)}, \ldots, \beta^{(i_\ell)} \sim f_{\beta(\lambda)|y}; \\ &\text{compute:} && \tfrac{1}{\ell} \textstyle\sum_{k=1}^{\ell} g(\beta^{(i_k)}), \end{aligned}$$

where ℓ is the number of Monte Carlo samples. Depending on the inferential tool, carrying out the first step above can be very expensive.

In the spirit of the present paper, so long as the target posterior arises from the required normal prior, as in (15), we consider instead using early stopped Langevin dynamics on the likelihood in the first step above. Following directly from (6), we propose the numerical integration scheme appearing below:

$$\begin{aligned} &\text{draw:} && \beta^{(k)} = \beta^{(k-1)} + \tau_k \cdot \nabla \log f_{y|\beta(\lambda)}(\beta^{(k-1)}) + \sqrt{2\tau_k} \cdot z^{(k)},\ k = k^*, \ldots, k^* + \ell; \\ &\text{compute:} && \tfrac{1}{\ell} \textstyle\sum_{k=k^*}^{k^*+\ell} g(\beta^{(k)}), \end{aligned} \tag{19}$$

where τ_k, $k = k^*, \ldots, k^* + \ell$, is a sequence of carefully chosen step sizes. In words, the above scheme runs the early-stopped Langevin iteration until it has converged to the target posterior, draws samples near the posterior mode, and then uses the associated sample average to estimate (18). It is critical to work in discrete time; therefore, the choice of step sizes is important, so that the early-stopped process does not diverge from the target posterior. We may use a fixed step size up until the stopping time $k^* = \lceil 1/(\tau\lambda) \rceil$, followed by step lengths that are square summable but not summable afterwards, e.g.,

$$\tau_j = \begin{cases} \tau, & \text{for } j \leq k^*, \text{ where } \tau < 1/L \\ \tau/j, & \text{for } j = k^* + 1, \ldots, k^* + \ell. \end{cases} \tag{20}$$

The following result, recalling the constants that were defined in Corollary 3, bounds the mean squared error of the above integration scheme, assuming fixed step sizes. We will abbreviate $\beta^{(1:k)} = (\beta^{(1)}, \ldots, \beta^{(k)})$.

Theorem 3 (Monte Carlo integration mean squared error)**.** *Assume the same conditions as in Corollary 3. Also, assume that $g : \mathbb{R}^p \mapsto \mathbb{R}$ is Q-Lipschitz continuous. Fix some number of iterations $k \geq k^*$, and a step size $\tau < 1/L$. Then, the early-stopped Monte Carlo integration scheme* (19)*, under the initialization $\beta^{(0)} \sim \text{Normal}(0, I_p/\lambda)$ satisfies the following inequalities.*

- *Under conditions A1 and A2:*

$$\begin{aligned} &\mathbb{E}_{\beta(\lambda), y, \beta^{(1:k)}} \left[\left(\frac{1}{k} \sum_{j=1}^{k} g(\beta^{(j)}) - \mathbb{E}_{\beta(\lambda)|y} \big[g(\beta(\lambda)) \big] \right)^2 \right] \\ &\qquad \leq 2Q^2 O(d^{3/2} \tau^{1/2}) + \frac{4Q^2}{m'_{R,\lambda} k} \sum_{j=1}^{k} \bigg(a_\lambda \cdot \exp(-m'_\lambda t_j) \\ &\qquad + \frac{b\lambda^2}{2} \cdot \frac{1 - \exp(-m'_\lambda t_j)}{m'_\lambda} + \frac{b_\lambda \lambda^2}{2} \cdot \frac{\exp\big((m'_\lambda - 2m)t\big) - 1}{m'_\lambda - 2m} \cdot \exp(-m'_\lambda t_j) \bigg), \end{aligned} \tag{21}$$

  for any $t_j \in [j\tau, (j+1)\tau)$.

- *Under conditions A1, A3, and A4:*

$$\begin{aligned} &\mathbb{E}_{\beta(\lambda), y, \beta^{(1:k)}} \left[\left(\frac{1}{k} \sum_{j=1}^{k} g(\beta^{(j)}) - \mathbb{E}_{\beta(\lambda)|y} \big[g(\beta(\lambda)) \big] \right)^2 \right] \\ &\qquad \leq 2Q^2 O(d^{3/2} \tau^{1/2}) + \frac{4Q^2}{m'_{R,\lambda} k} \sum_{j=1}^{k} \bigg(a_\lambda \cdot \exp(-m'_{R,\lambda} t_j) + \frac{b_R \lambda^2}{2} \cdot \frac{1 - \exp(-m'_{R,\lambda} t)}{m'_{R,\lambda}} \\ &\qquad + \frac{b_{R,\lambda} \lambda^2}{2} \cdot \frac{\exp\big((m'_{R,\lambda} - 2\alpha)t_j\big) - 1}{m'_{R,\lambda} - 2\alpha} \cdot \exp(-m'_{R,\lambda} t_j) \bigg), \end{aligned} \tag{22}$$

for any $t_j \in [j\tau, (j+1)\tau)$.

For technical reasons, we assume fixed step sizes in Theorem 3. However, we investigate diminishing step sizes, as in (20), through our experimental work to come in the next section. The result follows by relating the criterion (21) to the 1-Wasserstein distance via its dual representation, then using a comparison inequality to relate this distance to the 2-Wasserstein distance, and finally invoking Corollaries 3 and 4 appearing earlier in order to obtain the result.

7 Numerical simulations

We present several numerical examples supporting our general theory. Throughout, we focus on comparing the standard and early-stopped Langevin iterations, i.e., (5) and (6), by measuring the relative sampling efficiency, defined in (9). Importantly, we also check the tightness of the bounds we presented earlier, i.e., those from Sections 4, 5, and 6. We begin our presentation of the numerics by revisiting, and expanding on, the Gaussian data example that was first presented in Section 2, before turning to other Bayesian generalized linear models. Then, in line with Corollary 3 presented above, we show the results of an experiment involving a non-strongly-log-concave likelihood. Finally, we turn to experiments checking the accuracy of the early-stopped Monte Carlo integration scheme in (19).

7.1 Bayesian generalized linear models

7.1.1 Relative sampling efficiency

Gaussian regression. We start with the canonical Gaussian setup from Sections 2 and 4. For these experiments, we generate $\beta(\lambda)$ and $y \mid \beta(\lambda)$ by following the generative model in (3), for different values of the prior precision strength $\lambda \in \{0.1, 0.3, 0.5\}$ and a fixed noise variance $\sigma^2 = 1$. To form the data matrix, X, we first draw two random orthogonal matrices $U \in \mathbb{R}^{n\times p}$ and $V \in \mathbb{R}^{n\times p}$, and set the diagonal entries of $S \in \mathbb{R}^{p\times p}$ as $S_{ii} = \sqrt{|z_i|}$ with $z_i \overset{\text{iid}}{\sim} \text{Normal}(0,1)$, for $i = 1, \ldots, p$. Then, we form $X = USV^\top$. Here and in what follows, we always set the sample size $n = 500$, and the ambient dimension $p = 10$. This construction implies X has full column rank, almost surely. We experiment with both well and ill-conditioned data matrices (with condition numbers 4 and 40, respectively), by scaling maximum value of S suitably.

Now we simulate the stochastic differential equations (10), (11), corresponding to the standard and early-stopped Langevin dynamics, respectively, and compute the relative sampling efficiency (9) in closed form, using formulas detailed in the proofs in the Appendix. We plot the relative sampling efficiency vs. time t, for different draws of the singular values, in Figure 2. We also show the stopping time $t^* = 1/\lambda$, for each value of λ, with a dot in the figure. Throughout, our plots show the average of 25 trials. We can see that the relative sampling efficiency spikes near t^* and is small otherwise, just as our theory predicts (and as in Figure 1).

Logistic regression. We also consider a Bayesian logistic regression setup, where we generate:

$$\beta(\lambda) \sim \text{Normal}(0, I_p/\lambda), \quad \text{and} \quad y_i \mid \beta(\lambda) \sim \text{Bernoulli}(q_i), \text{ where } q_i = \frac{1}{1+\exp(-x_i^\top \beta)}.$$

To ensure the likelihood is strongly log-concave, we multiply it by (another) normal distribution having mean zero and covariance I_p/γ, $\gamma > 0$. Equivalently, we assume that $\beta(\lambda)$ follows a normal distribution with covariance

$$(1/\lambda + 1/\gamma) \cdot I_p,$$

where γ is small, e.g., $\gamma = 0.01$. Note that the likelihood is strongly log-concave, but not a density, which makes it an interesting test case. In this case, there is no closed-form expression for the

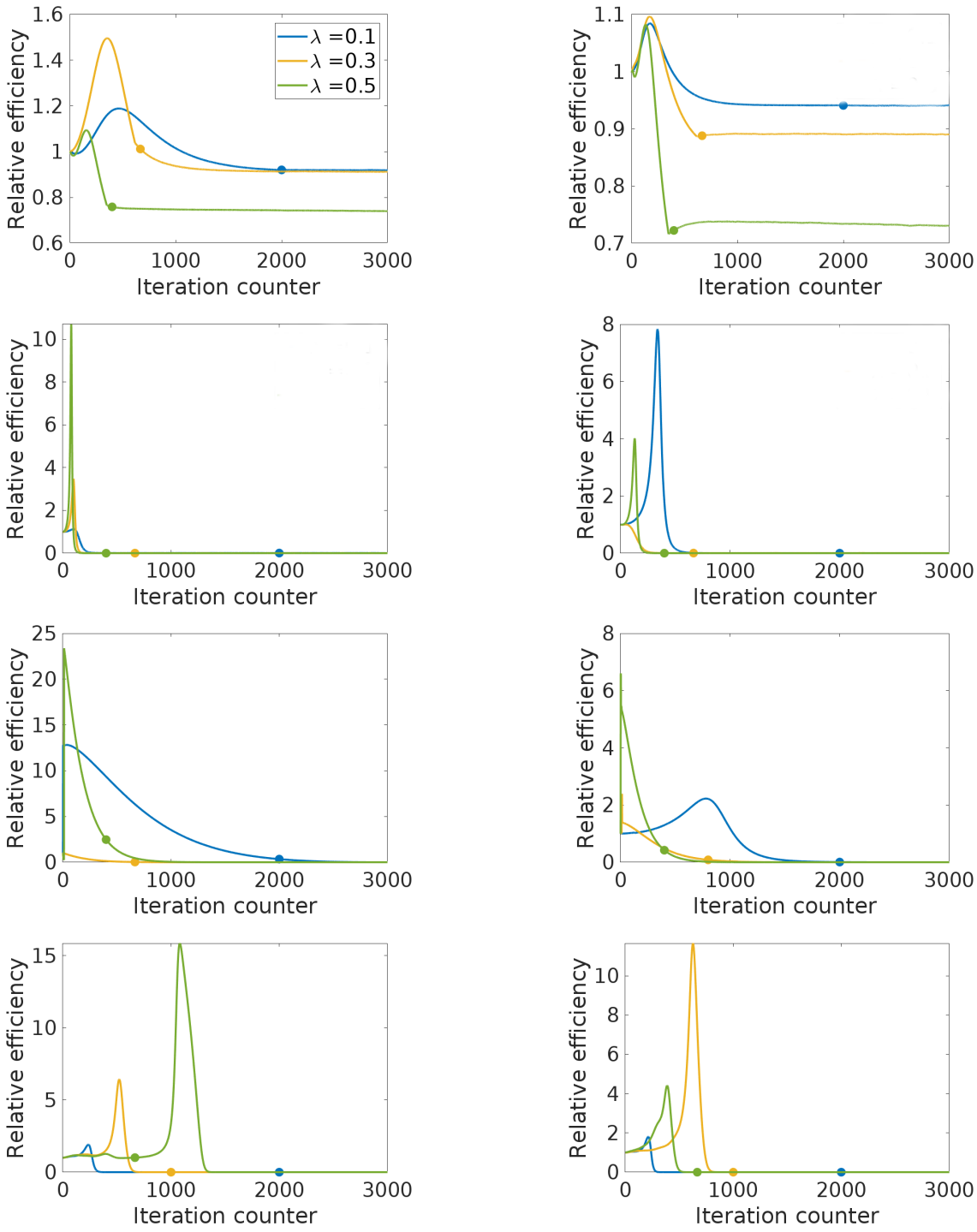

Figure 2: *The relative sampling efficiency, defined in* (9)*, of the early-stopped Langevin Monte Carlo dynamics* (6) *over the standard Langevin Monte Carlo dynamics* (5)*, for Bayesian linear regression (top row), logistic regression (second row), Poisson regression (third row), and a non-strongly-log-concave posterior (fourth row). The plots show the efficiencies for various prior precision strengths* $\lambda \in \{0.1, 0.3, 0.5\}$*, as well as for a well-conditioned data matrix* X *(left column) and an ill-conditioned one (right column). Throughout, the efficiency is large when the time* $t \approx 1/\lambda$*, and small otherwise.*

Wasserstein distance, so we compute it numerically using a linear time version of the Sinkhorn algorithm (Sinkhorn and Knopp, 1967; Cuturi, 2013; Peyré and Cuturi, 2019) due to Altschuler et al. (2017), i.e., we compute

$$W_2(\beta_{t^*},\, \beta(\lambda) \mid y) = \inf\Big\{\mathrm{tr}(P^\top C) : P \in \mathbb{R}_+^{\ell\times\ell},\; P\mathbb{1} = r,\; P^\top \mathbb{1} = c\Big\},$$

where $r \in \mathbb{R}^\ell, c \in \mathbb{R}^\ell$ are given vectors with positive entries that sum to one and $C \in \mathbb{R}_+^{\ell\times\ell}$ is a suitable cost matrix. Here each element of r and c is $1/l$, where l is the number of samples and the i,jth element of C stores the euclidean distance between ith and jth sample. The second row of Figure 2 shows the relative sampling efficiencies, where we see the same trend as in the Gaussian case.

Poisson regression. We again follow a setup similar to what was done above, except here we have:

$$y_i \mid \beta(\lambda) \sim \text{Categorical}(q_1, \ldots, q_\ell), \text{ where } q_j = \frac{\exp\big(y_i x_i^\top \beta - \exp(x_i^\top \beta)\big)}{y_i!}, \text{ for } j = 1, \ldots, \ell.$$

The third row of Figure 2 shows the corresponding results, where we can see the same pattern as mentioned above. As a whole, these plots reveal that the laws of β_t and $\beta(\lambda)$ are, in fact, close at the optimal stopping time t^*, and that this behavior is seemingly robust across different data models.

7.1.2 Tightness of bounds

In Figures 3 and 4 we investigate the tightness of our bounds, i.e., Theorem 1, Corollary 3, and Theorem 2, respectively, for the three data models above (Gaussian, logistic, and Poisson regression). To put all the problems on the same scale, we plot the relative log error in Figure 4,

$$\Big|\frac{\log \hat{W} - \log W^*}{\log W^*}\Big|, \tag{23}$$

where $\hat{W}$ denotes the numerical value of either of our bounds, and

$$W^* = \mathbb{E}_{\beta(\lambda),y}\big[W_2^2(\beta_{t^*},\, \beta(\lambda) \mid y)\big].$$

Figure 3 shows the bound presented in Theorem 1, for various values of λ, in the same Bayesian linear regression setup that was described above. It is clear that the bound is extremely sharp.

Turning to the general case, we plot both Corollary 3 and Theorem 2 in Figure 4. Here, we see that Theorem 2 is in fact sharper than Corollary 3; notably, for Poisson regression, it appears to be a few orders of magnitude sharper. Having said that, both our bounds do a good job on the Gaussian and logistic problems, where we know the Lipschitz constant L of the log likelihood, required to compute our bounds. However, note that the Poisson regression log likelihood is not globally Lipschitz smooth, only when restricted to compact sets. Therefore, for Poisson regression, we use as a numerical estimate of L the maximum spectral norm of the Hessian of the log likelihood computed over ten pilot simulations, which explains some of the apparent looseness of the bounds.

7.2 Non-strongly-log-concave posterior

We now look at the case when the likelihood function is m-strongly log-concave outside a ball of radius R, but not necessarily convex otherwise. Fix some R and a small constant $\epsilon > 0$ (in our experiments, we set $R = 4$ and $\epsilon = 0.01$). Now consider the negative log likelihood given, for $\beta \in \mathbb{R}$, by

$$F_{y|\beta}(\beta; y) = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\beta^2 - R^2 + 1\right)\sin^2(y_i)\right]\mathbb{1}_{\beta^2 \geq R^2} + \epsilon\beta^2 \tag{24}$$

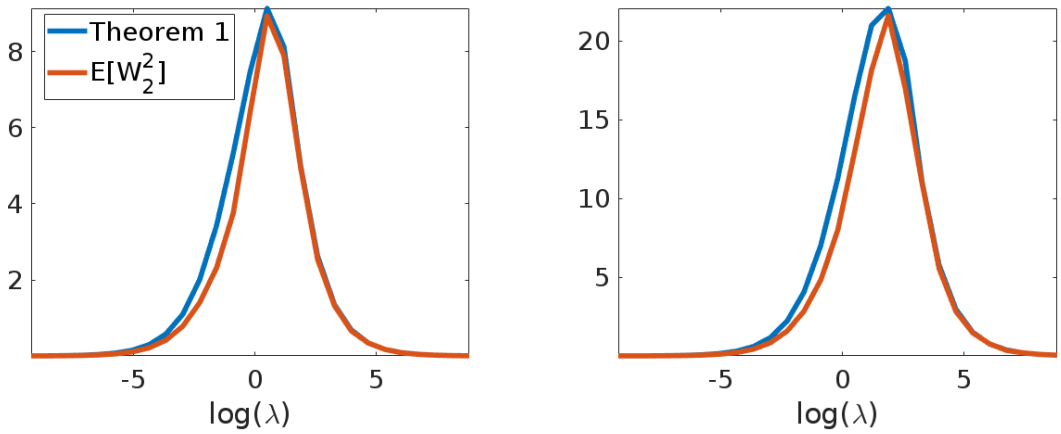

Figure 3: *The bound from Theorem 1 (blue) vs. the actual value (red) of the squared 2-Wasserstein distance between the law of early-stopped Langevin dynamics* (13) *at the optimal stopping time* $t^* = 1/\lambda$ *and the target posterior* $\beta(\lambda) \mid y$*, for Bayesian linear regression* (4)*. The plots show the values of the bound and ground truth, for various* λ*, as well as for a well-conditioned data matrix* X *(left) and an ill-conditioned one (right). The bound given by Theorem 1 is always very sharp.*

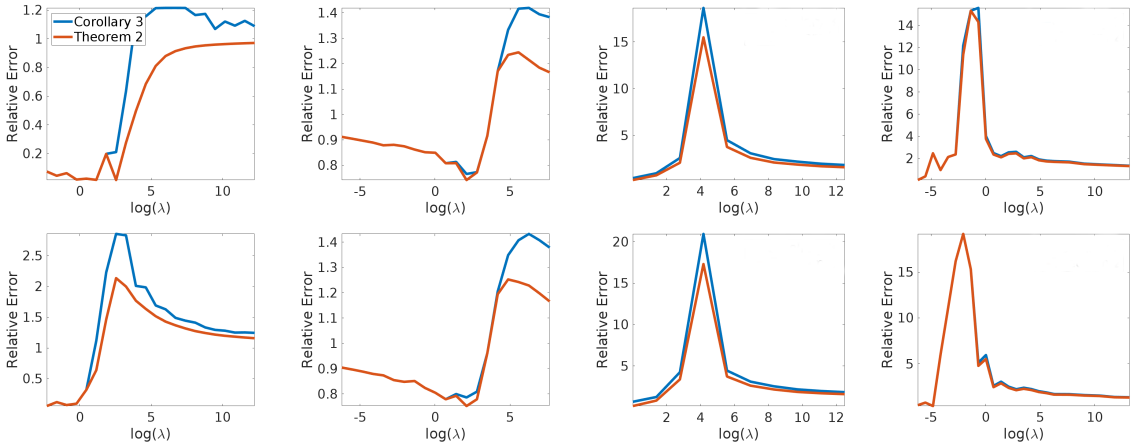

Figure 4: *The relative log error of the bounds from Corollary 3 (blue) and Theorem 2 (red), for Bayesian linear regression (first column), logistic regression (second column), Poisson regression (third column), and a non-strongly-log-concave posterior (fourth column). The plots show the values of the two bounds for various* λ*, as well as for a well-conditioned data matrix* X *(top row) and an ill-conditioned one (bottom row).*

$$+ \left[\left(\sin\left(\beta^2 - R^2\right) + \cos\left(\beta^2 - R^2\right)\right)\sin^2\left(y_i\right)\right] \mathbb{1}_{\beta^2 < R^2}.$$

The gradient and Hessian of $F_{y|\beta}(\beta; y)$ with respect to β are, almost surely with respect to Lebesgue measure on $\mathbb{R}$,

$$\begin{aligned}\nabla F_{y|\beta}(\beta; y) = \frac{1}{n}\sum_{i=1}^{n} &\left[2\beta\sin^2\left(y_i\right)\right] \mathbb{1}_{\beta^2 \geq R^2} + 2\epsilon\beta \\ &+ \left[\left(2\beta\cos\left(\beta^2 - R^2\right) - 2\beta\sin\left(\beta^2 - R^2\right)\right)\sin^2\left(y_i\right)\right] \mathbb{1}_{\beta^2 < R^2},\end{aligned}$$

and

$$\begin{aligned}\nabla^2 F_{y|\beta}(\beta; y) = \frac{1}{n}\sum_{i=1}^{n} &\left[2\sin^2\left(y_i\right)\right] \mathbb{1}_{\beta^2 \geq R^2} + 2\epsilon + \left[\left(2\cos\left(\beta^2 - R^2\right) - 2\sin\left(\beta^2 - R^2\right)\right)\sin^2\left(y_i\right)\right] \mathbb{1}_{\beta^2 < R^2} \\ &- \left[\left(4\beta^2\sin\left(\beta^2 - R^2\right) + 4\beta^2\cos\left(\beta^2 - R^2\right)\right)\sin^2\left(y_i\right)\right] \mathbb{1}_{\beta^2 < R^2}.\end{aligned}$$

Hence $F_{y|\beta}(\beta; y)$ is $(2 + 8\epsilon R^2)$-Lipschitz smooth, and satisfies Assumption A1. When $\beta^2 \geq R^2$, the negative log likelihood is strongly convex with parameter 2ϵ, so it also satisfies Assumption A3, but it is not convex in the region $\{\beta : \beta^2 < R^2\}$. From Proposition 1 in Ma et al. (2019b), we get the function satisfies Assumption A4 with $\alpha \geq \epsilon \exp(-16(2 + 8\epsilon R^2)R^2)$. See Figure 5. Now to draw samples from the associated posterior, we simply draw $\beta \sim \text{Normal}(0, 1/\lambda)$ from the prior, then draw y from the likelihood, which is proportional to $\exp(-F_{y|\beta}(\beta; y))$.

The bottom row of Figure 2 shows the relative sampling efficiencies, and the last column of Figure 4 investigates the tightness of Corollary 3 and Theorem 2; the results are roughly similar to those for Poisson regression, which is encouraging.

7.3 Quadrature

Finally, we investigate the accuracy of the early-stopped quadrature scheme, described in (19). We consider the Euclidean norm $z \mapsto \|z\|_2$ as our test function. Figure 6 shows the Monte Carlo integration error (21) vs. the number of iterations k. We can clearly see that the early-stopped scheme outperforms the standard Langevin-based scheme. The results for this test function are nearly identical to those for the squared Euclidean norm, i.e., $z \mapsto \|z\|_2^2$ (not shown), which is of course not Lipschitz continuous, and thus goes beyond the assumptions covered by Theorem 3.

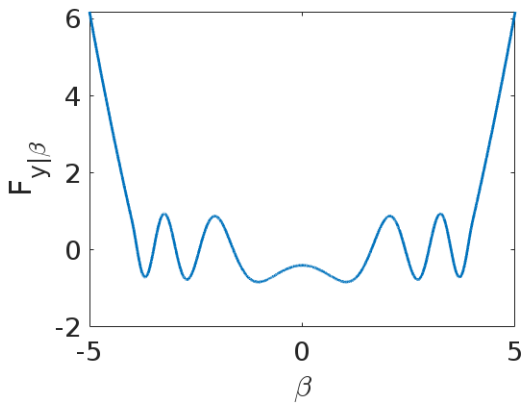

Figure 5: *The negative log likelihood from* (24).

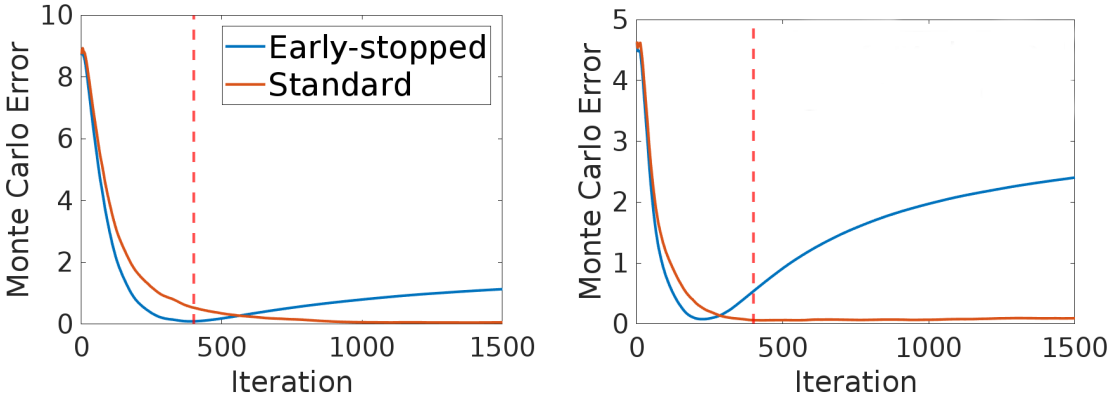

Figure 6: *The Monte Carlo integration error* (21) *vs. the number of iterations taken by the early-stopped numerical integration scheme* (19)*, for Bayesian linear regression* (4) *with the prior precision strength* $\lambda = 0.1$*. Here, we use the Euclidean norm as our test function. The left panel shows the results for a well-conditioned design, and the right shows the results for an ill-conditioned one.*

8 Conclusion

An exciting line of work, going back to Jordan et al. (1998) and continuing on recently through Dalalyan (2017b,a); Ma et al. (2019b); Wibisono (2018); Ma et al. (2019a); Cheng et al. (2019, 2020); Mou et al. (2021), has uncovered a number of intriguing connections between sampling algorithms and classical methods for optimization. These connections are useful, because they imply we can move ideas back and forth, i.e., from optimization to sampling, and from sampling back to optimization. As we see it though, implicit regularization has not really benefited yet from this sort of cross-pollination, and remains curiously underexplored. In the current paper, we sought to port ideas from implicit regularization to sampling, and carefully studied an early-stopped variant of the popular standard Langevin Monte Carlo iteration. We gave theoretical and empirical backing to the idea that an early-stopped variant of the usual Langevin Monte Carlo iteration can converge to a target posterior faster than the standard Langevin iteration.

There are a few directions to pursue as part of future work that would be interesting. First, it is clear that understanding the relationships between other, non-normal, priors and algorithms would be useful. A related question we might ask that seems interesting is: does putting a prior on the scale (i.e., not just the location, as we have done here) correspond to any particular algorithm? Examining and gaining a deeper understanding of the statistical properties—say, the risk—of the early-stopped Langevin iterates, i.e., going beyond sampling and numerical integration, certainly also appears within reach. More broadly though, we hope our paper motivates the continued exploration, and application, of ideas from implicit regularization to sampling.

9 Acknowledgements

We are indebted to Yi-An Ma for suggesting the topic of this paper to us. We also thank Ryan J. Tibshirani for encouraging discussions during the early stages of this project, and Xiang Cheng for a few helpful discussions. ED was partially supported by an NSF DMS CAREER award (2046874) and the NSF-Simons Collaboration on the Mathematical and Scientific Foundations of Deep Learning THEORINET (NSF 2031985).

10 Appendix

10.1 Proofs for Section 4

Our general strategy for proving the results in this section is (as mentioned in the main paper) to exploit the fact that in the canonical Gaussian setup from (3), both the linear stochastic processes (10), (11), as well as the posterior (4), have closed-form expressions that can simplify many of the arguments. Key to these arguments is the following result from Vatiwutipong and Phewchean (2019), collecting together several useful properties of a special kind of linear stochastic process, the Ornstein-Uhlenbeck process, which we invoke frequently.

Lemma 2 (Useful properties of the Ornstein-Uhlenbeck process (Vatiwutipong and Phewchean, 2019))**.** *Let $(\beta_t) \in \mathbb{R}^p$ be a p-dimensional Gaussian process satisfying*

$$d\beta_t = M[\mu - \beta_t]dt + \Gamma dW_t, \tag{25}$$

where $M, \Gamma \in \mathbb{S}^p_{++}$ and $\mu \in \mathbb{R}^p$, so that the process (β_t) is an Ornstein-Uhlenbeck process. Then, for $t \in [0, T]$ with $T > 0$, the unique solution to (25) is given by

$$\beta_t = \exp(-Mt)\beta_0 + [I_p - \exp(-Mt)]\mu + \int_0^t \exp\big((s-t)M\big)\Gamma dW_s, \tag{26}$$

which has the following properties:

- $\mathbb{E}(\beta_t) = \exp(-Mt)\beta_0 + [I_p - \exp(-Mt)]\mu$;
- $\mathrm{Cov}(\beta_t) = \int_0^t \exp(M(s-t))\Gamma\Gamma^\top \exp(M^\top(s-t))ds$;
- $\mathrm{vec}(\mathrm{Cov}(\beta_t)) = (M \oplus M)^+[I_p - \exp((-M \oplus M)t)]\mathrm{vec}(\Gamma\Gamma^\top)$.

Here, $A \oplus B$ denotes the Kronecker sum of A and B, $\mathrm{vec}(A)$ denotes the usual vectorization of the matrix A, A^+ denotes the Moore-Penrose pseudo-inverse of A, and $\exp(A)$ denotes the matrix exponential of A.

Now we carry on with proving the results of Section 4.

10.1.1 Proof of Theorem 1

Noting that in the Gaussian setting the early-stopped process (13) is in fact a Gaussian process of the form (25), we may apply Lemma 2 above with $M = X^\top X$ and $\mu = (X^\top X)^+ X^\top y$, giving

$$\beta_t \sim \mathrm{Normal}\bigg([I_p - \exp(-tX^\top X)]X^+ y,\ (X^\top X)^+[I_p - \exp(-2tX^\top X)]\bigg), \tag{27}$$

conditional on $\beta(\lambda), y$.

Now it is straightforward to check, using an eigendecomposition, that the covariance matrices associated with the early-stopped process, as in (27), and the posterior, as in (4), commute; therefore, we may invoke a well-known result (see, e.g., Proposition 7 in Givens and Shortt (1984), or Takatsu (2011)) characterizing the squared 2-Wasserstein distance between two normals,

$$\begin{aligned} W_2^2(\beta_t,\, \beta(\lambda)) = &\underbrace{\Big\|[(I_p - \exp(-tX^\top X))(X^\top X)^+ - (X^\top X + \lambda I)^{-1}]X^\top y\Big\|_2^2}_{T_1} \\ &+ \underbrace{\Big\| \left[(X^\top X)^+(I_p - \exp(-2tX^\top X))\right]^{1/2} - (X^\top X + \lambda I)^{-1/2}\Big\|_F^2}_{T_2}, \end{aligned} \tag{28}$$

again conditional on $\beta(\lambda), y$.

As a result, we can see that bounding the quantity of interest, i.e., $\mathbb{E}_{\beta(\lambda),y}[W_2^2(\beta_{t^*},\, \beta(\lambda) \mid y)]$, boils down to tightly controlling the two terms associated with the means and covariances above, i.e., T_1, T_2, respectively. The helper Lemmas 3, 4, appearing below, give the required control, and show that

$$\mathbb{E}_{\beta(\lambda),y}\big[W_2^2(\beta_t,\, \beta(\lambda) \mid y)\big] = \sum_{i=1}^{p} \frac{h(s_i; t, \lambda)}{s_i}, \tag{29}$$

where $s_1, \ldots, s_p$ denote the singular values of $X^\top X$, and we defined the function

$$h(x; t, \lambda) = \left[\sqrt{1 - \exp(-2tx)} - \sqrt{\frac{x}{x+\lambda}}\right]^2 + (x/\lambda + 1)\left[\frac{\lambda}{x+\lambda} - \exp(-tx)\right]^2.$$

We actually prove a somewhat more general result here than the one stated in the theorem, which reduces to the stated result as a special case. Here, instead of simply plugging the parametrization $t = 1/\lambda$ into (29) and simplifying, we consider a more flexible parametrization, where $t = c/\lambda$, for some fixed $c > 0$. We seek the value of c making (29) as small as possible; simply setting $c = 1$ recovers the result given in the main paper.

Therefore, substituting the parametrization $t = c/\lambda$, and making a simple change of variables, in (29), yields

$$\begin{aligned}
h(x; t, c/t) &= \left[\sqrt{1 - \exp(-2tx)} - \sqrt{\frac{x}{x + c/t}}\right]^2 + (xt/c + 1)\left[\frac{c/t}{x + c/t} - \exp(-tx)\right]^2 \\
&= \left[\sqrt{1 - \exp(-2tx)} - \sqrt{\frac{tx}{tx + c}}\right]^2 + (tx/c + 1)\left[\frac{c}{tx + c} - \exp(-tx)\right]^2 \\
&= \left[\sqrt{1 - \exp(-2z)} - \sqrt{\frac{z}{z + c}}\right]^2 + (z/c + 1)\left[\frac{c}{z + c} - \exp(-z)\right]^2 := H(z; c).
\end{aligned}$$

We aim to solve a variational problem for each summand, i.e., we want to find c such that A, B are small in the expression

$$H(z; c) \le \frac{Az \wedge Bc}{z + c}.$$

Now write $I(z; c) = (z + c)H(z; c)$, so that

$$I(z; c) = \left\{\sqrt{[1 - \exp(-2z)] \cdot (z + c)} - \sqrt{z}\right\}^2 + \left[c - (z + c)\exp(-z)\right]^2 / c.$$

In Figure 7, we plot $I(z; c)/c$ with $c = 1$, i.e., $I(z; 1)$. We can see that the optimal $A := A_c, B := B_c$ must satisfy

$$\begin{aligned}
A_c &\ge \sup_{z \ge 0} I(z; c)/z, \\
B_c &\ge \sup_{z \ge 0} I(z; c)/c.
\end{aligned}$$

In principle, the strategy we pursued above works for all $c > 0$. However, we may simply put $c = 1$, and numerically maximize both $z \mapsto I(z; 1)/z$, and $z \mapsto I(z; 1)/c$, showing that $A = A_1$ is at most 0.2379, and that B is upper bounded by 1.026, respectively, which are evidently sharp enough for our purposes. Therefore, we have shown that

$$\mathbb{E}_{\beta(\lambda),y}\big[W_2^2(\beta_t,\, \beta(\lambda) \mid y)\big] \le \sum_{i=1}^{p} \frac{As_i \wedge B/t}{s_i(s_i + 1/t)},$$

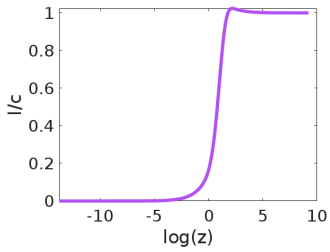

Figure 7: *The function $I(z;1)$, used in the proof of Theorem 1.*

and since we assumed $\lambda = c/t$, with $c = 1$, the claimed result follows. □

We make a couple of additional remarks on the result.

- In our proof of the result above, we assumed the initialization $\beta_0 = 0$, but another fairly natural initialization is $\beta_0 \sim \text{Normal}(0, I_p/\lambda)$; here, we show that the same result still goes through, when working with this alternative initialization, just with minor modifications.

  Let $X = US^{1/2}V^\top$ be a singular value decomposition. Then, looking back at (29), we can calculate, conditional on $\beta(\lambda), y$, that

$$\begin{aligned} T_1 &= \Big\| \exp(-tX^\top X)\beta(\lambda) + [(I_p - \exp(-tX^\top X))(X^\top X)^+ - (X^\top X + \lambda I)^{-1}]X^\top y \Big\|_2^2 \\ &= \Big\| V \exp(-tS)V^\top \beta(\lambda) + VMU^\top y \Big\|_2^2 \\ &= \Big\| \exp(-tS)V^\top \beta(\lambda) + MU^\top y \Big\|_2^2 \\ &= \sum_{i=1}^p (\exp(-ts_i)v_i^\top \beta(\lambda) + M_{ii}u_i^\top y)^2. \end{aligned}$$

  Here, $M = M(S, \lambda, t)$ denotes a diagonal matrix depending on S, λ, t, identified above. Therefore, expanding the square, we have that

$$T_1 = \sum_{i=1}^p |u_i^\top y|^2 \left[\frac{1 - \exp(-ts_i)}{\sqrt{s_i}} - \frac{\sqrt{s_i}}{s_i + \lambda}\right]^2 + |v_i^\top \beta(\lambda)|^2 \exp(-2ts_i) + 2u_i^\top y \cdot v_i^\top \beta(\lambda) \cdot C_i,$$

  for some constants C_i, $i = 1, \ldots, p$, that do not depend on $\beta(\lambda)$ or y. Taking expectations over $\beta(\lambda), y$ shows that the cross terms vanish, i.e., that we only need to add to $\mathbb{E}_{\beta(\lambda),y}(T_1)$ the terms

$$\mathbb{E}_{\beta(\lambda),y}\big[|v_i^\top \beta(\lambda)|^2 \exp(-2ts_i)\big] = \exp(-2ts_i)/\lambda, \quad i = 1, \ldots, p.$$

- We also mention that writing $X^+ = (X^\top X)^+ X^\top y$ in (27) draws a connection to Example 1 studied in Wibisono (2018).

10.1.2 Statement and proof of helper Lemma 3

Lemma 3. *Let $X = US^{1/2}V^\top$ be a singular value decomposition. Fix $\lambda > 0$. Then, the term T_1 from (28) satisfies*

$$\mathbb{E}_{\beta(\lambda),y}(T_1) = \sum_{i=1}^p (s_i/\lambda + 1)\left[\frac{1 - \exp(-ts_i)}{\sqrt{s_i}} - \frac{\sqrt{s_i}}{s_i + \lambda}\right]^2.$$

Proof. For notational convenience, we define

$$\begin{aligned} A &= (I_p - \exp(-tX^\top X))(X^\top X)^+ X^\top \\ &= V(I_p - \exp(-tS))S^+ S^{1/2} U^\top \\ B &= (X^\top X + \lambda I)^{-1} X^\top = V(S + \lambda I)^+ S^{1/2} U^\top. \end{aligned}$$

Then, we may simply calculate, conditional on $\beta(\lambda), y$, that

$$\begin{aligned} \|Ay - By\|_2^2 &= y^\top A^\top A y - 2y^\top A^\top B y + y^\top B^\top B y \\ &= y^\top U S^{1/2} S^+ (I_p - \exp(-tS))(I_p - \exp(-tS)) S^+ S^{1/2} U^\top y \\ &+ y^\top U S^{1/2} (S + \lambda I_p)^{-1} (S + \lambda I_p)^+ S^{1/2} U^\top y \\ &- 2y^\top U S^{1/2} S^+ (I_p - \exp(-tS))(S + \lambda I_p)^+ S^{1/2} U^\top y \\ &= \sum_{i=1}^p |u_i^\top y|^2 \left[\frac{1 - \exp(-ts_i)}{\sqrt{s_i}} - \frac{\sqrt{s_i}}{s_i + \lambda} \right]^2. \end{aligned}$$

Now let $\varepsilon \sim \text{Normal}(0, \sigma^2 \cdot I_n)$. Taking expectations over the randomness in $\beta(\lambda), y$ we see that

$$\begin{aligned} u_i^\top \mathbb{E}_{\beta(\lambda),\varepsilon}(yy^\top) u_i &= u_i^\top \mathbb{E}_{\beta(\lambda),\varepsilon}[(X\beta(\lambda) + \varepsilon)(X\beta(\lambda) + \varepsilon)^\top] u_i \\ &= u_i^\top \mathbb{E}_{\beta(\lambda),\varepsilon}[X\beta(\lambda)\beta(\lambda)^\top X^\top + \varepsilon\varepsilon^\top] u_i \\ &= u_i^\top \mathbb{E}_{\beta(\lambda),\varepsilon}[XX^\top/\lambda + I_p] u_i \\ &= s_i/\lambda + 1, \end{aligned}$$

i.e.,

$$\mathbb{E}_{\beta(\lambda),y}(T_1) = \sum_{i=1}^p (s_i/\lambda + 1) \left[\frac{1 - \exp(-ts_i)}{\sqrt{s_i}} - \frac{\sqrt{s_i}}{s_i + \lambda} \right]^2,$$

as claimed. □

10.1.3 Statement and proof of helper Lemma 4

Lemma 4. *Let $X = US^{1/2}V^\top$ be a singular value decomposition. Fix $\lambda > 0$. Then, the term T_2 from* (28) *satisfies*

$$\mathbb{E}_{\beta(\lambda),y}(T_2) = \sum_{i=1}^p \left[\sqrt{\frac{1 - \exp(-2ts_i)}{s_i}} - \sqrt{\frac{1}{s_i + \lambda}} \right]^2.$$

Proof. The result follows via calculations similar to those found in the proof of Lemma 3. First, we define

$$\begin{aligned} A &= (X^\top X)^+ (I_p - \exp(-2tX^\top X)) \\ &= VS^+ (I_p - \exp(-2tS))V^\top, \end{aligned}$$

and

$$B = (X^\top X + \lambda I)^{-1} = V(S + \lambda I)^+ V^\top.$$

Then, we simply calculate

$$\|A^{1/2} - B^{1/2}\|_F^2 = \text{tr}(A^{1/2} - B^{1/2})(A^{1/2} - B^{1/2})]$$

$$= \mathrm{tr}[A + B - 2(AB)^{1/2}]$$
$$= \sum_{i=1}^{p} \left[\sqrt{\frac{1 - \exp(-2ts_i)}{s_i}} - \sqrt{\frac{1}{s_i + \lambda}} \right]^2,$$

which gives the result. □

10.1.4 Proof of Corollary 1

Our goal is to show the concentration of the Wasserstein distance with respect to the randomness due to $\beta(\lambda)$ and $\varepsilon \sim \mathrm{Normal}(0, \sigma^2 \cdot I_n)$. Recall from equation (28) that we have, for a constant c that does not depend on β, ε, that

$$W_2^2(\beta_t, \beta(\lambda)) = \|Qy\|_2^2 + c = R + c,$$
$$Q := [(I_p - \exp(-tX^\top X))(X^\top X)^+ - (X^\top X + \lambda I)^{-1}]X^\top.$$

This shows that R is a quadratic form of y. In more detail, we define $Z = [\lambda^{1/2}\beta; \varepsilon]$, and note that $Z \sim \mathcal{N}(0, I_{n+p})$, We have that

$$R = Z^\top M Z$$
$$M = A^\top A$$
$$A = Q[\lambda^{-1/2} X; I_n].$$

Due to the orthogonal invariance of the Gaussian distribution, i.e., $OZ \overset{d}{=} Z$ for all orthogonal matrices O, we can diagonalize the matrix M, and get that R has the distribution of a weighted sum of $\chi^2(1)$ random variables, where the weights are the eigenvalues of M. Let us define the eigenvalues of M to be $m_i \geq 0$, $i = 1, \ldots, n + p$. Then, we have

$$R \overset{d}{=} \sum_{i=1}^{n+p} m_i Z_i^2.$$

Thus, to derive the concentration of the Wasserstein distance, we need to know how weighted sums of chi-squared random variables concentrate. While this is not a standard topic in introductory probability or machine learning, there has been a great deal of work on understanding the distribution of such weighted chi-squared random variables. See, e.g., Robbins and Pitman (1949); Gurland (1953); Jensen and Solomon (1972); Davis (1977); Solomon and Stephens (1977); Gabler and Wolff (1987); Bausch (2013), and references therein. However, since here we are interested only in a concentration inequality for the Wasserstein distance, we can use the simpler Bernstein inequality for sub-exponential random variables, or equivalently the Hanson-Wright inequality for random matrices with sub-Gaussian entries. See Vershynin (2018) for a discussion of these topics. Bernstein's inequality (see, e.g., Theorem 2.8.1 in Vershynin (2018)), or equivalently the Hanson-Wright inequality (e.g., Theorem 6.2.1 in Vershynin (2018)), show that for some constant $c > 0$, and all $w \geq 0$,

$$P\{|Z^\top M Z - \mathrm{tr}M| \geq w\} \leq 2 \exp\left[-c \min\left(\frac{w^2}{\|M\|_F^2}, \frac{w}{\|M\|_{op}} \right) \right].$$

In our case, we can write that

$$\|M\|_F^2 = \|A\|_4^4$$
$$\|M\|_{op} = \|A\|_\infty^2,$$

where $\|A\|_q$ is the q-Schatten norm of A. This proves the claim.

10.1.5 Proof of Lemma 1

The proof of the result is conceptually similar to that of Theorem 1, so we skip some of the details. First of all, as noted earlier, because the standard process (10) is also of the form (25), we may once again invoke Lemma 2 from above, this time with $M = X^\top X + \lambda I_p$ and $\mu = (X^\top X + \lambda I_p)^{-1} X^\top y$. Doing so, we obtain

$$\tilde{\beta}_t \sim \mathcal{N}\bigg([I_p - \exp(-t[X^\top X + \lambda I_p])]\beta(\lambda),\, (X^\top X + \lambda I_p)^{-1}[I_p - \exp(-2t(X^\top X + \lambda I_p))] \bigg),$$

conditional on $\beta(\lambda), y$. Then, following steps similar to those found in the proof of Theorem 1, we can calculate for the standard process that

$$\begin{aligned} W_2^2(\tilde{\beta}_t,\, \beta(\lambda) \mid y) = &\Big\| \exp(-t[X^\top X + \lambda I_p])\beta(\lambda) \Big\|_2^2 \\ &+ \Big\| (X^\top X + \lambda I)^{-1/2} \Big\{ \big[I_p - \exp(-2t(X^\top X + \lambda I_p)) \big]^{1/2} - I_p \Big\} \Big\|_F^2, \end{aligned}$$

again conditional on $\beta(\lambda), y$. Taking expectations in the previous display gives

$$\mathbb{E}_{\beta(\lambda),y}\big[W_2^2(\tilde{\beta}_t,\, \beta(\lambda) \mid y) \big] = \sum_{i=1}^p \tilde{h}(s_i; t, \lambda),$$

where we write

$$\tilde{h}(s; t, \lambda) = \exp(-2t[s+\lambda]) \frac{s/\lambda + 1}{(s+\lambda)^2} + \frac{\Big[\big\{ 1 - \exp(-2t[s+\lambda]) \big\}^{1/2} - 1 \Big]^2}{s+\lambda},$$

in analogy to the quantity $h(x; t, \lambda)$ defined in the proof of Theorem 1.

To find the Wasserstein distance at the stopping time $t^* = 1/\lambda$, we may plug the stopping time into the previous display and simplify, as in

$$\begin{aligned} \tilde{h}(s_i; 1/\lambda, \lambda) &= \exp(-2[s/\lambda + 1]) \frac{1}{\lambda(s+\lambda)} + \frac{\Big[\big\{ 1 - \exp(-2[s/\lambda + 1]) \big\}^{1/2} - 1 \Big]^2}{s+\lambda} \\ &:= \tilde{h}(s, \lambda). \end{aligned}$$

To obtain the claimed result, it is enough to show that $\tilde{h}(s, \lambda) \geq U(s, \lambda)$, where $U(s, \lambda) = \frac{As \wedge B\lambda}{s(s+\lambda)}$ is a summand in our upper bound on the Wasserstein distance between the early-stopped process β_t and the posterior $\beta(\lambda) \mid y$, from Theorem 1. But this is equivalent to showing, for $x = s/\lambda$, that

$$\frac{\exp(-2[x+1])/\lambda + \Big[\big\{ 1 - \exp(-2[x+1]) \big\}^{1/2} - 1 \Big]^2}{A \wedge (B/x)} \geq 1.$$

Suppose that x is such that $A \wedge (B/x) = A$, i.e., $A \leq B/x$, or that $x \leq B/A$. Then, it is enough to show that in the regime of interest, the bias term is large enough, namely that

$$\exp(-2[x+1])/\lambda \geq A.$$

Rearranging, we find that for each fixed $x \geq 0$, it is enough that

$$\lambda \leq A^{-1} \exp(-2[x+1]).$$

This shows that, for each fixed value of $x = s/\lambda \leq B/A$, if λ is small enough, then the early-stopped process at time $t = 1/\lambda$ improves over early stopping with the standard process at the same time, completing the proof.

10.1.6 Proof of Corollary 4

Key to our proof of the result is the following helper lemma, which is a modest extension to Theorem 1 from Cheng et al. (2020). In its original form, Theorem 1 in Cheng et al. (2020) bounds the discretization error so long as the process noise is itself bounded, which is obviously not the case in (6). Below is our version of the result, which accommodates Gaussian noise.

Lemma 5 (Extension to Theorem 1 in Cheng et al. (2020))**.** *Let $\beta^{(k)}$ and β_t have dynamics as defined in* (6) *and* (11) *respectively, $t \in [k\tau, (k+1)\tau)$ and the initial condition satisfy $\mathbb{E}\left[\|\beta^{(0)}\|_2^2\right] \leq \frac{p}{\lambda}$ and $\mathbb{E}\left[\|\beta_0\|_2^2\right] \leq \frac{p}{\lambda}$. Define the constants*

$$L_N = \frac{4\gamma L_\xi}{c_m},\ \alpha_q = \frac{L + L_N^2}{2c_m^2},\ R_q = max\left\{R, \frac{16\gamma^2 L_N}{mc_m}\right\},$$
$$\eta = \min\left\{\frac{m}{2}, \frac{2c_m^2}{32R_q^2}\right\} \exp\left(-\frac{7}{3}\alpha_q R_q^2\right),\ 0 < c < \sqrt{2},\ \gamma = \sqrt{p}$$

and L_ξ is small positive constant. Let $\hat{\epsilon} \leq \left(\frac{16(L+L_N^2)}{\eta}\right) \exp(7\alpha_q R_q/3) \frac{R_q}{\alpha_q R_q^2 + 1}$. Let τ be a step size satisfying

$$\tau \leq \min\left\{\frac{\frac{\eta^2\hat{\epsilon}^2}{512\gamma(L^2+L_N^4)\exp\left(\frac{14\alpha_q}{R_q^2}\right)}}{\frac{2\eta\hat{\epsilon}}{\gamma(L^2+L_N^4)\exp\left(\frac{7\alpha_q}{R_q^2}\right)\sqrt{p/\lambda}}}\right.\!.$$

If we assume that $\beta^{(0)} = \beta_0$, then there exists a coupling between $\beta^{(k)}$ and β_t such that

$$\mathbb{E}\left[\|\beta^{(k)} - \beta_t\|_2\right] \leq \hat{\epsilon}.$$

With Lemma 5 in hand, proving Corollary 4 is immediate. As a result of the helper lemma, we have $W_1(\beta^{(k)}, \beta_t) = O(p^{1/2}\tau^{1/2})$. Since $W_2^2(\beta^{(k)}, \beta_t) \leq p \cdot W_1(\beta^{(k)}, \beta_t)$, we have $W_2^2(\beta^{(k)}, \beta_t) = O(p^{3/2}\tau^{1/2})$. For more details, readers can refer to Cheng et al. (2020).

10.2 Proofs for Section 5

In this section, we start by stating and proving a general result (i.e., Theorem 4, appearing below), from which Corollaries 2 and 3 in the main paper follow. Then, we show how the corollaries may be obtained from the general result. We also prove the sharper (but less transparent) bound, i.e., Theorem 2 in the main paper, relating the Kullback-Leibler divergence between the distribution of the early-stopped process β_t and the posterior $\beta(\lambda) \mid y$.

The arguments given in this section differ from those given in the previous section, as in the general case, there is no convenient expression for the law of the early-stopped process, or for the posterior distribution. Therefore, we proceed somewhat indirectly here. Our jumping off point is the classical theory of gradient flows on function spaces (Ambrosio et al., 2008), which gives us some control over the Kullback-Leibler divergence between the law of the early-stopped process and the posterior; the main technical challenges that arise are due to the fact that the stationary distribution of the early-stopped process β_t does *not* coincide with the target posterior $f_{\beta(\lambda)|y}$ — a peculiar (and challenging) feature of our problem setting.

Below is the main result of this section, followed by its proof. For all of the proofs in this section, we omit the dependence on λ, R in the subscripts for any of the defined constants (unlike in the actual statements of the theorems), in order to keep the notation below light.

Theorem 4 (Bound on expected K-L divergence, (non-)strongly log-concave posterior)**.** *Assume the data model* (15)*, as well as condition A1. Additionally, assume either (i) condition A2, or (ii)*

conditions A3 and A4, i.e., for some $R \geq 0$, assume that the likelihood $f_{y|\beta(\lambda)}$ is $m'_{R,\lambda}$-strongly log-concave outside of a ball centered around zero with radius R. Finally, assume there exist constants $b_1, b_2, q > 0$, such that for all R and $s \geq 0$, we have the uniform bound

$$\mathbb{E}_{\beta_s} \|\beta_s\|_2^2 \leq b_1 + b_2 \exp(-qs). \tag{30}$$

Now define the constants

$$m'_{R,\lambda} = \begin{cases} m + \lambda, & R = 0, \\ \lambda, & R > 0, \end{cases}$$

and $c = \lambda^2/2$. Then, there exist constants $a_1, a_2 > 0$, such that the law of early-stopped Langevin dynamics (11), at time $t > 0$ and under the initialization $\beta_0 \sim \mathrm{Normal}(0, I_p/\lambda)$, satisfies

$$\mathbb{E}_{\beta(\lambda),y}\big[D_{\mathrm{kl}}(\beta_t \| \beta(\lambda) \mid y)\big] \leq a_1 \cdot \exp\left(-m'_{R,\lambda} t\right) + b_1 c \cdot \frac{1 - \exp(-m'_{R,\lambda} t)}{m'_{R,\lambda}}$$
$$+ b_2 c \cdot \frac{\exp((m'_{R,\lambda} - q)t) - 1}{m'_{R,\lambda} - q} \cdot \exp(-m'_{R,\lambda} t).$$

Proof. Recall that, under the data model (15), the target posterior has the form

$$f_{\beta(\lambda)|y}(\beta; y) \propto \exp\Big(- F_{y|\beta(\lambda)}(\beta; y) - (\lambda/2) \cdot \|\beta\|_2^2 \Big). \tag{31}$$

Write f_{β_t} for the Radon-Nikodym derivative of $\mathrm{Law}(\beta_t)$, i.e., the density of the early-stopped process at time t. Then, the probability measure $\mathrm{Law}(\beta_t)$, and the velocity field $\nabla f_{y|\beta(\lambda)} + \nabla \log f_{\beta_t}$, satisfy the usual continuity (Fokker-Planck) equation. Therefore, Lemma 1 in Cheng and Bartlett (2018) applies, itself stemming from the classical theory of gradient flows (Ambrosio et al., 2008), giving

$$\frac{dD_{\mathrm{kl}}(\beta_t \| \beta(\lambda) \mid y)}{dt} = -\mathbb{E}_{\beta_t}\left[\left\langle \nabla \log\left(\frac{f_{\beta_t}(\beta_t)}{f_{\beta(\lambda)|y}(\beta_t)}\right), \nabla \log f_{\beta_t}(\beta_t) + \nabla F_{y|\beta(\lambda)}(\beta_t) \right\rangle\right],$$

conditional on $\beta(\lambda), y$. Adding and subtracting $\nabla \log f_{\beta(\lambda)|y}$, then expanding, shows this equals

$$-\mathbb{E}_{\beta_t}\left[\left\langle \nabla \log\left(\frac{f_{\beta_t}(\beta_t)}{f_{\beta(\lambda)|y}(\beta_t)}\right), \nabla \log\left(\frac{f_{\beta_t}(\beta_t)}{f_{\beta(\lambda)|y}(\beta_t)}\right) + \left(\nabla \log f_{\beta(\lambda)|y}(\beta_t) + \nabla F_{y|\beta(\lambda)}(\beta_t)\right) \right\rangle\right]$$
$$= -\mathbb{E}_{\beta_t}\left\| \nabla \log\left(\frac{f_{\beta_t}(\beta_t)}{f_{\beta(\lambda)|y}(\beta_t)}\right)\right\|_2^2 - \mathbb{E}_{\beta_t}\left[\left\langle \nabla \log\left(\frac{f_{\beta_t}(\beta_t)}{f_{\beta(\lambda)|y}(\beta_t)}\right), \nabla \log f_{\beta(\lambda)|y}(\beta_t) + \nabla F_{y|\beta(\lambda)}(\beta_t) \right\rangle\right].$$

Plugging (31) into the previous display, and using the basic inequality $a^\top b \leq (1/2)(\|a\|_2^2 + \|b\|_2^2)$, we get

$$\begin{aligned} \frac{dD_{\mathrm{kl}}(\beta_t \| \beta(\lambda) \mid y)}{dt} &= -\mathbb{E}_{\beta_t}\left\| \nabla \log\left(\frac{f_{\beta_t}(\beta_t)}{f_{\beta(\lambda)|y}(\beta_t)}\right)\right\|_2^2 + \mathbb{E}_{\beta_t}\left[\left\langle \nabla \log\left(\frac{f_{\beta_t}(\beta_t)}{f_{\beta(\lambda)|y}(\beta_t)}\right), \lambda \cdot \beta_t \right\rangle\right] \\ &\leq -\mathbb{E}_{\beta_t}\left\| \nabla \log\left(\frac{f_{\beta_t}(\beta_t)}{f_{\beta(\lambda)|y}(\beta_t)}\right)\right\|_2^2 + \frac{1}{2} \cdot \mathbb{E}_{\beta_t}\left\| \nabla \log\left(\frac{f_{\beta_t}(\beta_t)}{f_{\beta(\lambda)|y}(\beta_t)}\right)\right\|_2^2 + \frac{\lambda^2}{2} \cdot \mathbb{E}_{\beta_t}\|\beta_t\|_2^2 \\ &\leq -\frac{1}{2} \cdot \mathbb{E}_{\beta_t}\left\| \nabla \log\left(\frac{f_{\beta_t}(\beta_t)}{f_{\beta(\lambda)|y}(\beta_t)}\right)\right\|_2^2 + \frac{\lambda^2}{2} \cdot \mathbb{E}_{\beta_t}\|\beta_t\|_2^2. \end{aligned} \tag{32}$$

Now we use the fact that the posterior $f_{\beta(\lambda)|y}$ satisfies Assumption A2 (or Assumptions A3 and A4), i.e., that it is m'-strongly log-concave (possibly outside of a certain ball), which gives the bound

$$\frac{dD_{\mathrm{kl}}(\beta_t \| \beta(\lambda) \mid y)}{dt} \leq -m' \cdot D_{\mathrm{kl}}(\beta_t \| \beta(\lambda) \mid y) + \frac{\lambda^2}{2} \cdot \mathbb{E}_{\beta_t} \|\beta_t\|_2^2. \tag{33}$$

An application of Gronwall's inequality (see, e.g., Lakshmikantham et al. (1989)) gives

$$D_{\mathrm{kl}}(\beta_t \| \beta(\lambda) \mid y) \leq D_{\mathrm{kl}}(\beta_0 \| \beta(\lambda) \mid y) \cdot \exp\left(-m't\right) + c \cdot \int_0^t \mathbb{E}_{\beta_s} \|\beta_s\|_2^2 \cdot \exp\left(-m'[t-s]\right) ds. \tag{34}$$

Now we must control the integral appearing on the right-hand side of (34), above. But from (30), we have

$$\int_0^t \mathbb{E}_{\beta_s} \|\beta_s\|_2^2 \cdot \exp\left(-m'[t-s]\right) ds \leq \int_0^t \left(b_1 + b_2 \exp(-qs)\right) \exp\left(-m'[t-s]\right) ds,$$

and using the identity $\int_0^t \exp(bs) ds = [\exp(bt) - 1]/b$, we can explicitly calculate that

$$\begin{aligned}
&\int_0^t \mathbb{E}_{\beta_s} \|\beta_s\|_2^2 \exp\left(-m'[t-s]\right) ds \leq b_1 \left(\frac{1 - \exp(-m't)}{m'}\right) + \int_0^t b_2 \exp(-qs - m't + m's) ds \\
&= b_1 \left(\frac{1 - \exp(-m't)}{m'}\right) + b_2 \left(\frac{\exp((m'-q)t) - 1}{m' - q}\right) \exp(-m't).
\end{aligned} \tag{35}$$

Finally, substituting the bound on the integral, in (35), back into the bound on the Kullback-Leibler divergence, in (34), gives

$$D_{\mathrm{kl}}(\beta_t \| \beta(\lambda) \mid y) \leq a_1 \left(\exp\left(-m't\right) + b_1 c\right) \left(\frac{1 - \exp(-m't)}{m'}\right) + b_2 c \left(\frac{\exp((m'-q)t) - 1}{m' - q}\right) \exp(-m't). \tag{36}$$

Here, we wrote $a_1 \geq D_{\mathrm{kl}}(\beta_0 \| \beta(\lambda) \mid y)$, for another numerical constant $a_1 > 0$. Taking expectations over $\beta(\lambda), y$ completes the proof. □

10.2.1 Proof of Corollary 2

Under the conditions assumed by the corollary, the helper Lemmas 6 and 8 appearing below give us control over $\mathbb{E}_{\beta_t} \|\beta_t\|_2^2$, as well as the Kullback-Leibler divergence between the initial point $\beta_0 \sim \mathrm{Normal}(0, I_p/\lambda)$ and the posterior, required to apply Theorem 4. Putting these lemmas together with Theorem 4, we see

$$\begin{aligned}
\mathbb{E}_{\beta(\lambda),y}\big[D_{\mathrm{kl}}(\beta_t \| \beta(\lambda) \mid y)\big] &\leq \left(\frac{pL}{2\lambda} + \frac{p}{2} \log \frac{\lambda}{m'}\right) \exp\left(-m't\right) \\
&+ 2c \left(\frac{p}{m} + \|\mu\|_2^2\right) \left(\frac{1 - \exp(-m't)}{m'}\right) + 2c \frac{p}{m} \left(\frac{L - \lambda}{\lambda} + \log \frac{\lambda}{m}\right) \left(\frac{\exp((\lambda - m)t) - 1}{\lambda - m}\right) \exp(-m't).
\end{aligned}$$

Making the appropriate definitions gives the result.

10.2.2 Statement and proof of helper Lemma 6

Lemma 6. *Assume the same conditions, and constants, as in Corollary 2. Then, for any $t > 0$, it follows that*

$$\mathbb{E}_{\beta_t} \|\beta_t\|_2^2 \leq 2 \left(\frac{p}{m} + \|\mu\|_2^2\right) + \frac{2p}{m} \left(\frac{L}{\lambda} - 1 + \log \frac{\lambda}{m}\right) \cdot \exp(-2mt).$$

Proof. Let π denote the stationary distribution of the early-stopped process. We choose an auxiliary random variable $\beta_\infty \sim \pi \propto \exp(-f_{\beta_\infty})$ with $\mu = \mathbb{E}(\beta_\infty)$ which couples optimally with β_t: $(\beta_t, \beta_\infty) \sim P \in \mathcal{P}(\mathrm{Law}(\beta_t), \mathrm{Law}(\beta_\infty))$. Then, adding and subtracting β_∞ shows

$$\begin{aligned}\mathbb{E}_{\beta_t}\|\beta_t\|_2^2 &= \mathbb{E}_{(\beta_t,\beta_\infty)\sim P}\Big[\|\beta_\infty + (\beta_t - \beta_\infty)\|_2^2\Big] \\ &\leq 2\cdot \mathbb{E}_{\beta_\infty\sim\pi}\|\beta_\infty\|_2^2 + 2\cdot \mathbb{E}_{(\beta_t,\beta_\infty)\sim P}\Big[\|\beta_t - \beta_\infty\|_2^2\Big]. \end{aligned} \tag{37}$$

Now there are (at least) two ways to bound the first term on the right-hand side of (37), $\mathbb{E}_{\beta_\infty\sim\pi}\|\beta_\infty\|_2^2$. The first is to simply use part (ii) of Proposition 1 in Durmus and Moulines (2019), in order to get

$$\mathbb{E}_{\beta_\infty\sim\pi}\|\beta_\infty\|_2^2 \leq \frac{p}{m} + \|\mu\|_2^2.$$

The second way is to invoke the Brascamp-Lieb inequality (Brascamp and Lieb, 1976a,b). Concretely, one can take the p coordinate functions x_i, $i = 1, \ldots, p$, and sum the resulting bounds

$$\mathrm{Var}\big((\beta_\infty)_i\big) \leq \mathbb{E}_{\beta_\infty\sim\pi}\Big[\big(\nabla^2 f_{\beta_\infty}(\beta_\infty)\big)^{-1}_{ii}\Big],$$

to get

$$\mathbb{E}_{\beta_\infty\sim\pi}\Big[\|\beta_\infty - \mu\|_2^2\Big] \leq \mathbb{E}_{\beta_\infty\sim\pi}\Big[\mathrm{tr}\big(\nabla^2 f_{\beta_\infty}(\beta_\infty)\big)^{-1}\Big].$$

Then, using that $\nabla^2 f_{\beta_\infty}(\beta) \succeq m \cdot I_p$, for all $\beta \in \mathbb{R}^p$, we have

$$\mathbb{E}_{\beta_\infty\sim\pi}\Big[\mathrm{tr}\big(\nabla^2 f_{\beta_\infty}(\beta_\infty)\big)^{-1}\Big] \leq p/m,$$

giving the required bound.

To bound the second term on the right-hand side in (37), we would like to exploit the transportation cost inequality (16). Therefore, we put together the well-known contraction inequality (see, e.g., Chapter 2 in Villani (2008))

$$W_2(\beta_t, \pi) \leq \exp(-mt)\cdot W_2(\beta_0, \pi),$$

with (16) and the helper Lemma 7 appearing below to get, for the second term above, that

$$\begin{aligned}\mathbb{E}_{(\beta_t,\beta_\infty)\sim P}\Big[\|\beta_t - \beta_\infty\|_2^2\Big] \leq W_2^2(\beta_t, \pi) &\leq \exp(-2mt)\cdot W_2^2(\beta_0, \pi) \\ &\leq \exp(-2mt)\cdot \frac{2}{m}\cdot D_{\mathrm{kl}}(\beta_0\|\pi) \\ &\leq \exp(-2mt)\cdot \frac{2p}{m}\left(\frac{L}{\lambda} - 1 + \log\frac{\lambda}{m}\right).\end{aligned}$$

Putting together the two bounds completes the proof. □

10.2.3 Statement and proof of helper Lemma 7

Lemma 7. *Assume the same conditions, and constants, as in Corollary 2. Then, for any $t > 0$, it follows that*

$$D_{kl}(\beta_0\|\pi) \leq \frac{p(L-\lambda)}{2\lambda} + \frac{p}{2}\log\frac{\lambda}{m}.$$

In the proofs of the remaining helper lemmas in this section, we reuse some of the notation from the proof of Lemma 6.

Proof. Let

$$\beta^* \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} f_{\beta_\infty}(\beta; y).$$

Also, for any $\beta \in \mathbb{R}^p$, let

$$\bar{f}(\beta) = f_{\beta_\infty}(\beta; y) - f_{\beta_\infty}(\beta^*; y).$$

$$\begin{aligned} D_{\mathrm{kl}}(\beta_0 \| \pi) &= \int f_{\beta_0}(\beta) \log \left(\frac{f_{\beta_0}(\beta)}{\pi(\beta)} \right) d\beta \\ &= \int f_{\beta_0}(\beta) \log f_{\beta_0}(\beta) d\beta - \int f_{\beta_0}(\beta) \log \pi(\beta) d\beta. \end{aligned} \tag{38}$$

So, it is enough to bound the two terms on the right-hand side of (38) above, separately.

We start with the second term. Observe, for all $\beta \in \mathbb{R}^p$, that we have both

$$\pi(\beta) = \frac{\exp(-\bar{f}(\beta))}{\int \exp(-\bar{f}(\beta')) d\beta'},$$

and

$$\frac{m}{2} \|\beta\|_2^2 \leq \bar{f}(\beta) \leq \frac{L}{2} \|\beta\|_2^2,$$

with the latter pair of inequalities following by our Assumptions A1 and A2. Therefore, we have that

$$\begin{aligned} -\log \pi(\beta) &= \bar{f}(\beta) + \log \int \exp\left(-\bar{f}(\beta')\right) d\beta' \\ &\leq \frac{L}{2} \|\beta\|_2^2 + \log \int \exp\left(-\frac{m}{2} \|\beta'\|_2^2 \right) d\beta' \\ &= \frac{L}{2} \|\beta\|_2^2 + \frac{p}{2} \log \frac{2\pi}{m}, \end{aligned}$$

implying that

$$\begin{aligned} -\int f_{\beta_0}(\beta) \log \pi(\beta) d\beta &\leq \int \left(\frac{L}{2} \|\beta\|_2^2 + \frac{p}{2} \log \frac{2\pi}{m} \right) \left(\frac{\lambda}{2\pi} \right)^{\frac{p}{2}} \exp\left(\frac{-\lambda}{2} \|\beta\|_2^2 \right) d\beta \\ &\leq \frac{pL}{2\lambda} + \frac{p}{2} \log \frac{2\pi}{m}. \end{aligned}$$

In a similar way, we can calculate for the first term in (38) that

$$\int f_{\beta_0}(\beta) \log f_{\beta_0}(\beta) d\beta = -\frac{p}{2} \log \frac{2\pi}{\lambda} - \frac{p}{2}.$$

Putting it all together, we have that

$$\begin{aligned} D_{\mathrm{kl}}(\beta_0 \| \pi) &\leq \frac{pL}{2\lambda} + \frac{p}{2} \log \frac{2\pi}{m} - \frac{p}{2} \log \frac{2\pi}{\lambda} - \frac{p}{2} \\ &= \frac{p(L-\lambda)}{2\lambda} + \frac{p}{2} \log \frac{\lambda}{m}, \end{aligned}$$

as required. □

10.2.4 Proof of Corollary 3

Similar to what was done in the proof of Corollary 2, we use helper Lemmas 8 and 9, which are applicable under the assumed conditions, to get

$$\begin{aligned}\mathbb{E}_{\beta(\lambda),y}\big[D_{\mathrm{kl}}(\beta_t\|\beta(\lambda)\mid y)\big] &\leq \frac{pL}{2\lambda}\exp\left(-\lambda t\right) + b_1 c\left(\frac{1-\exp(-\lambda t)}{\lambda}\right) \\ &\quad + cb_2\left(\frac{\exp((\lambda-\alpha)t)-1}{\lambda-\alpha}\right)\exp(-\lambda t),\end{aligned}$$

where

$$\begin{aligned}b_1 &= 2\left(\frac{8p}{\alpha}\log\frac{2L}{m} + \frac{256}{\alpha}\frac{L^2}{m^2}LR^2 + \|\mu\|_2^2\right), \\ b_2 &= \frac{2}{\alpha}\left(\frac{p(L-\lambda)}{2\lambda} + \frac{p}{2}\log\frac{2\lambda}{m} + 32\frac{L^2}{m^2}LR^2\right).\end{aligned}$$

The result follows after making the appropriate definitions.

10.2.5 Statement and proof of helper Lemma 8

Lemma 8. *Assume the same conditions, and constants, as in Corollary 3. Then, for any $t > 0$, it follows that*

$$D_{kl}(\beta_0\|\beta(\lambda)\mid y) \leq \begin{cases}\frac{pL}{2\lambda} + \frac{p}{2}\log\frac{\lambda}{m+\lambda}, & R = 0, \\ \frac{pL}{2\lambda}, & R > 0.\end{cases}$$

Proof. The proof is identical to that of Lemma 7, with the following two modifications. First, when $R = 0$, we set m (in Lemma 7) to $m + \lambda$. And second, when $R > 0$, we set m (in Lemma 7) to λ, and L (in Lemma 7) to $L + \lambda$. □

10.2.6 Statement and proof of helper Lemma 9

Lemma 9. *Assume the same conditions, and constants, as in Corollary 3. Then, for any $t > 0$, it follows that*

$$\mathbb{E}_{\beta_t}\|\beta_t\|_2^2 \leq b_1 + b_2\exp(-2\alpha t),$$

where

$$\begin{aligned}b_1 &= 2\left(\frac{8p}{\alpha}\log\frac{2L}{m} + \frac{256}{\alpha}\frac{L^2}{m^2}LR^2 + \|\mu\|_2^2\right), \\ b_2 &= \frac{2}{\alpha}\left(\frac{p(L-\lambda)}{2\lambda} + \frac{p}{2}\log\frac{2\lambda}{m} + 32\frac{L^2}{m^2}LR^2\right).\end{aligned}$$

Proof. The proof is similar to that of Lemma 6. Just as was done in that proof, some manipulations show

$$\begin{aligned}\mathbb{E}_{\beta_t}\|\beta_t\|_2^2 &= \mathbb{E}_{(\beta_t,\beta_\infty)\sim P}\Big[\|\beta_\infty + (\beta_t - \beta_\infty)\|_2^2\Big] \\ &\leq 2\cdot\mathbb{E}_{\beta_\infty\sim\pi}\|\beta_\infty\|_2^2 + 2\cdot\mathbb{E}_{(\beta_t,\beta_\infty)\sim P}\Big[\|\beta_t - \beta_\infty\|_2^2\Big] \\ &\leq 2\left(\frac{8p}{\alpha}\log\frac{2L}{m} + \frac{256}{\alpha}\frac{L^2}{m^2}LR^2 + \|\mu\|_2^2\right) + 2W_2^2(\beta_t,\pi). \qquad (39)\end{aligned}$$

Here, to get (39), we used the inequality

$$\mathbb{E}_{\beta_\infty \sim \pi} \|\beta_\infty\|_2^2 \leq \frac{8p}{\alpha} \log \frac{2L}{m} + \frac{256}{\alpha} \frac{L^2}{m^2} LR^2 + \|\mu\|_2^2,$$

which follows from Lemma 5 in Ma et al. (2019b).

On the other hand, using (16) and helper Lemma 10, we have

$$\begin{aligned} W_2^2(\beta_t, \pi) &\leq \frac{2}{\alpha} \cdot D_{\mathrm{kl}}(\beta_t \| \pi) \\ &\leq \frac{2}{\alpha} \exp(-2\alpha t) \cdot D_{\mathrm{kl}}(\beta_0 \| \pi) \\ &\leq \frac{2}{\alpha} \left(\frac{p(L-\lambda)}{2\lambda} + \frac{p}{2} \log \frac{2\lambda}{m} + 32 \frac{L^2}{m^2} LR^2 \right). \end{aligned} \tag{40}$$

Plugging (40) into (39) gives

$$\begin{aligned} \mathbb{E}\|\beta_t\|_2^2 \leq 2 \left(\frac{8p}{\alpha} \log \frac{2L}{m} + \frac{256}{\alpha} \frac{L^2}{m^2} LR^2 + \|\mu\|_2^2 \right) + \\ \frac{2}{\alpha} \left(\frac{p(L-\lambda)}{2\lambda} + \frac{p}{2} \log \frac{2\lambda}{m} + 32 \frac{L^2}{m^2} LR^2 \right) \cdot \exp(-2\alpha t), \end{aligned}$$

which shows the result. □

10.2.7 Statement and proof of helper Lemma 10

Lemma 10. *Assume the same conditions, and constants, as in Corollary 3. Then, for any $t > 0$, it follows that*

$$D_{kl}(\beta_0 \| \pi) \leq \frac{p(L-\lambda)}{2\lambda} + \frac{p}{2} \log \frac{2\lambda}{m} + 32 \frac{L^2}{m^2} LR^2.$$

Proof. The proof is similar to that of Lemma 7, so we only mention the major changes here. In the setting of the present lemma, we have the following lower bound from Ma et al. (2019b):

$$\bar{f}(\beta) \geq \frac{m}{4} \|\beta'\|_2^2 - 32 \frac{L^2}{m^2} LR^2.$$

Therefore, just as in the proof of Lemma 7, we can calculate

$$\begin{aligned} -\log \pi(\beta) &= \bar{f}(\beta) + \log \int \exp\left(-\bar{f}(\beta')\right) d\beta' \\ &\leq \frac{L}{2} \|\beta\|_2^2 + \log \int \exp\left(-\frac{m}{4} \|\beta'\|_2^2 + 32 \frac{L^2}{m^2} LR^2 \right) d\beta' \\ &= \frac{L}{2} \|\beta\|_2^2 + \frac{p}{2} \log \frac{4\pi}{m} + 32 \frac{L^2}{m^2} LR^2. \end{aligned}$$

This gives

$$\begin{aligned} -\int f_{\beta_0}(\beta) \log \pi(\beta) d\beta &\leq \int \log \pi(\beta) \left(\frac{\lambda}{2\pi} \right)^{\frac{p}{2}} \exp\left(\frac{-\lambda}{2} \|\beta\|_2^2 \right) d\beta \\ &\leq \frac{pL}{2\lambda} + \frac{p}{2} \log \frac{4\pi}{m} + 32 \frac{L^2}{m^2} LR^2. \end{aligned}$$

Therefore, we have

$$\begin{aligned} D_{\mathrm{kl}}(\beta_0\|\pi) &= \int f_{\beta_0}(\beta)\log f_{\beta_0}(\beta)d\beta - \int f_{\beta_0}(\beta)\log\pi(\beta)d\beta \\ &\leq \frac{pL}{2\lambda} + \frac{p}{2}\log\frac{4\pi}{m} + 32\frac{L^2}{m^2}LR^2 - \frac{p}{2}\log\frac{2\pi}{\lambda} - \frac{p}{2} \\ &= \frac{p(L-\lambda)}{2\lambda} + \frac{p}{2}\log\frac{2\lambda}{m} + 32\frac{L^2}{m^2}LR^2, \end{aligned}$$

which completes the proof. □

10.2.8 Proof of Theorem 2

The structure of the proof is similar to that of Theorem 4 above; therefore, we only highlight the main differences here. To start, we follow the same logic as in the proof of Theorem 4 through (32), but then invoke Young's inequality in its more general form, i.e., $2x^\top y \leq a\|x\|_2^2 + \|y\|_2^2/a$, for some $a > 0$. Doing so, and simplifying, gives

$$\begin{aligned} \frac{dD_{\mathrm{kl}}(\beta_t\|\beta(\lambda)\mid y)}{dt} &= -\,\mathbb{E}_{\beta_t}\left\|\nabla\log\left(\frac{f_{\beta_t}(\beta_t)}{f_{\beta(\lambda)|y}(\beta_t)}\right)\right\|_2^2 + \mathbb{E}_{\beta_t}\left[\left\langle\nabla\log\left(\frac{f_{\beta_t}(\beta_t)}{f_{\beta(\lambda)|y}(\beta_t)}\right),\lambda\beta_t\right\rangle\right] \\ &\leq -\,\mathbb{E}_{\beta_t}\left\|\nabla\log\left(\frac{f_{\beta_t}(\beta_t)}{f_{\beta(\lambda)|y}(\beta_t)}\right)\right\|_2^2 + \frac{a}{2}\cdot\mathbb{E}_{\beta_t}\left\|\nabla\log\left(\frac{f_{\beta_t}(\beta_t)}{f_{\beta(\lambda)|y}(\beta_t)}\right)\right\|_2^2 + \frac{\lambda^2}{2a}\cdot\mathbb{E}_{\beta_t}\|\beta_t\|_2^2 \\ &\leq (a/2-1)\cdot\mathbb{E}_{\beta_t}\left\|\nabla\log\left(\frac{f_{\beta_t}(\beta_t)}{f_{\beta(\lambda)|y}(\beta_t)}\right)\right\|_2^2 + \frac{\lambda^2}{2a}\cdot\mathbb{E}_{\beta_t}\|\beta_t\|_2^2, \end{aligned}$$

conditional on $\beta(\lambda), y$. Then, just as in (32), Assumption A2 (or Assumptions A3 and A4) imply

$$\frac{dD_{\mathrm{kl}}(\beta_t\|\beta(\lambda)\mid y)}{dt} \leq (a/2-1)m'\cdot D_{\mathrm{kl}}(\beta_t\|\beta(\lambda)\mid y) + c\cdot\mathbb{E}_{\beta_t}\|\beta_t\|_2^2,$$

where $a < 2$, and $c = \lambda^2/(2a)$. We then follow essentially the same steps through (36), with $m'' = 2(1-a/2)m'$, instead of m', giving

$$D_{\mathrm{kl}}(\beta_t\|\beta(\lambda)\mid y) \leq a_1\exp\left(-m''t\right) + b_1 c\left(\frac{1-\exp(-m''t)}{m''}\right) + b_2 c\left(\frac{\exp((m''-\alpha)t)-1}{m''-\alpha}\right)\exp(-m''t). \tag{41}$$

Now observe that the bound in (41) has the form

$$G(t) = q + r\exp(-ut) + v\exp(-wt),$$

where

$$\begin{aligned} q &= b_1\frac{c}{m''}, \qquad u = m'', \qquad w = \alpha, \\ r &= a_1 - b_1\frac{c}{m''} - b_2 c\frac{1}{m''-\alpha},\ v = b_2 c\frac{1}{m''-\alpha} \end{aligned}$$

are constants that do not depend on t. Therefore, differentiating $G(t)$ with respect to t gives

$$\begin{aligned} G'(t) &= r\exp(-ut)' + v\exp(-wt)' \\ &= -ru\exp(-ut) + -vw\exp(-wt) \\ &= [-ru\exp([w-u]t) + -vw]\exp(-wt), \end{aligned}$$

so that the optimal stopping time $t^* > 0$ must satisfy $ru\exp([w-u]t) + vw = 0$, i.e.,

$$t^* = \begin{cases} \frac{1}{w-u}\log\left(-\frac{wv}{ru}\right), & \text{if } \frac{wv}{ru} < 0 \\ \infty, & \text{otherwise.} \end{cases}$$

Putting it all together, we get that

$$D_{\mathrm{kl}}(\beta_{t^*} \| \beta(\lambda) \mid y) \leq \inf_{a\in(0,2)} G(a;\lambda,m,L,p),$$

and taking expectations completes the proof.

10.3 Proof of Theorem 3

We start by bounding $\mathbb{E}_{f_{\beta^{(1)}},\ldots,f_{\beta^{(k)}}}\left[\left(\frac{1}{k}\sum_{j=1}^k g(\beta^{(j)}) - \mathbb{E}_{f_{\beta(\lambda)|y}}\left[g(\beta(\lambda))\right]\right)^2\right]$ conditioned on y. For notational convenience, for the rest of this proof, let us define

$$g(y) = \mathbb{E}_{f_{\beta^{(1)}},\ldots,f_{\beta^{(k)}}}\left[\left(\frac{1}{k}\sum_{j=1}^k g(\beta^{(j)}) - \mathbb{E}_{f_{\beta(\lambda)|y}}\left[g(\beta(\lambda))\right]\right)^2\right].$$

Using Jensen's inequality, we can write

$$\begin{aligned} g(y) &\leq \frac{1}{k}\sum_{j=1}^k \mathbb{E}_{f_{\beta^{(1)}},\ldots,f_{\beta^{(k)}}}\left[\left(g(\beta^{(j)}) - \mathbb{E}_{f_{\beta(\lambda)|y}}\left[g(\beta(\lambda))\right]\right)^2\right] \\ &= \frac{1}{k}\sum_{j=1}^k \mathbb{E}_{f_{\beta^{(k)}}}\left[\left(g(\beta^{(j)})\right)^2 + \left(\mathbb{E}_{f_{\beta(\lambda)|y}}\left[g(\beta(\lambda))\right]\right)^2 - 2g(\beta^{(j)})\mathbb{E}_{f_{\beta(\lambda)|y}}\left[g(\beta(\lambda))\right]\right] \\ &\leq \frac{1}{k}\sum_{j=1}^k \mathbb{E}_{f_{\beta^{(k)}}}\left[\left(g(\beta^{(j)})\right)^2 + \mathbb{E}_{f_{\beta(\lambda)|y}}\left[\left(g(\beta(\lambda))\right)^2\right] - 2g(\beta^{(j)})\mathbb{E}_{f_{\beta(\lambda)|y}}\left[g(\beta(\lambda))\right]\right]. \end{aligned}$$

Given y is fixed, $\beta^{(j)}$ and $\beta(\lambda)|y$ are independent, we can then write the above inequality as

$$\begin{aligned} g(y) &\leq \frac{1}{k}\sum_{j=1}^k \mathbb{E}_{f_{\beta^{(k)}},f_{\beta(\lambda)|y}}\left[\left(g(\beta^{(j)}) - g(\beta(\lambda))\right)^2\right] \\ &\leq \frac{Q^2}{k}\sum_{j=1}^k \mathbb{E}\left[\|\beta^{(j)} - \beta(\lambda)\|^2\right]. \end{aligned}$$

Last inequality follows from Lipschitz continuity of g. Let $\beta_t \sim f_{\beta_t}$ and $\beta(\lambda) \sim f_{\beta(\lambda)|y}$ with an optimal coupling $(\beta_t, \beta(\lambda))$ so that $\mathbb{E}[\|\beta_t - \beta(\lambda)\|^2] = W_2(\beta_t, \beta(\lambda))$. Then

$$\begin{aligned} g(y) &\leq \frac{Q^2}{k}\sum_{j=1}^k \mathbb{E}\left[\|\beta^{(j)} - \beta_t + \beta_t - \beta(\lambda)\|_2^2\right] \\ &\leq \frac{2Q^2}{k}\sum_{j=1}^k \mathbb{E}\left[\|\beta^{(j)} - \beta_t\|_2^2 + \|\beta_t - \beta(\lambda)\|_2^2\right] \\ &= \frac{2Q^2}{k}\sum_{j=1}^k \left(\mathbb{E}\left[\|\beta^{(j)} - \beta_t\|_2^2 + W_2(\beta_t, \beta(\lambda))\right]\right). \end{aligned}$$

Let $\beta_t \sim f_{\beta_t}$ and $\beta^{(j)} \sim f_{\beta^{(j)}}$ with an optimal coupling $(\beta_t, \beta^{(j)})$ so that $\mathbb{E}[\|\beta_t - \beta(\lambda)\|_2^2] = W_2(\beta_t, \beta(\lambda))$. Then

$$g(y) \leq \frac{2Q^2}{k} \sum_{j=1}^{k} \left(W_2^2 \left(\beta^{(j)}, \beta_t \right) + W_2^2 \left(\beta_t, \beta(\lambda) \right) \right).$$

For $t_j \in [j\tau, (j+1)\tau)$, using Lemma 5, and equation 36 in proof of Theorem 4 along with Talagrand inequality, we have

$$\begin{aligned} g(y) \leq 2Q^2 O(d^{3/2}\tau^{1/2}) + \frac{4Q^2}{m'_{R,\lambda} k} \sum_{j=1}^{k} \big(& a_1 \cdot \exp\left(-m'_{R,\lambda} t_j\right) \\ & + b_1 c \cdot \frac{1 - \exp(-m'_{R,\lambda} t)}{m'_{R,\lambda}} + b_2 c \cdot \frac{\exp((m'_{R,\lambda} - q)t_j) - 1}{m'_{R,\lambda} - q} \cdot \exp(-m'_{R,\lambda} t_j) \big), \end{aligned}$$

where a_1, b_1 and b_2 are defined in Theorem 4, and $c = \frac{\lambda^2}{2}$. Substituting a_1, b_1 and b_2

- in the strongly convex case, we get

$$\begin{aligned} g(y) \leq 2Q^2 O(d^{3/2}\tau^{1/2}) + \frac{4Q^2}{m'_{R,\lambda} k} \sum_{j=1}^{k} \Bigg(& \left(\frac{pL}{2\lambda} + \frac{p}{2} \log \frac{\lambda}{m'} \right) \exp\left(-m' t\right) \\ & + 2c \left(\frac{p}{m} + \|\mu\|_2^2 \right) \left(\frac{1 - \exp(-m' t_j)}{m'} \right) + 2c \frac{p}{m} \left(\frac{L - \lambda}{\lambda} + \log \frac{\lambda}{m} \right) \left(\frac{\exp((\lambda - m)t_j) - 1}{\lambda - m} \right) \exp(-m' t_j) \Bigg), \end{aligned}$$

- in the non-strongly convex case, we get

$$\begin{aligned} g(y) \leq 2Q^2 O(d^{3/2}\tau^{1/2}) + \frac{4Q^2}{m'_{R,\lambda} k} \sum_{j=1}^{k} \Bigg(& \frac{pL}{2\lambda} \exp\left(-\lambda t_j\right) + b_1 c \left(\frac{1 - \exp(-\lambda t_j)}{\lambda} \right) \\ & + c b_2 \left(\frac{\exp((\lambda - \alpha)t_j) - 1}{\lambda - \alpha} \right) \exp(-\lambda t_j) \Bigg), \end{aligned}$$

where

$$\begin{aligned} b_1 &= 2 \left(\frac{8p}{\alpha} \log \frac{2L}{m} + \frac{256}{\alpha} \frac{L^2}{m^2} LR^2 + \|\mu\|_2^2 \right), \\ b_2 &= \frac{2}{\alpha} \left(\frac{p(L - \lambda)}{2\lambda} + \frac{p}{2} \log \frac{2\lambda}{m} + 32 \frac{L^2}{m^2} LR^2 \right). \end{aligned}$$

References

A. Ali, J. Z. Kolter, and R. J. Tibshirani. A continuous-time view of early stopping for least squares regression. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1370–1378. PMLR, 2019.

A. Ali, E. Dobriban, and R. Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *International Conference on Machine Learning*, pages 233–244. PMLR, 2020.

J. Altschuler, J. Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1961–1971, 2017.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

D. Bakry and M. Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84*, pages 177–206. Springer, 1985.

J. Bausch. On the efficient calculation of a linear combination of chi-square random variables with an application in counting string vacua. *Journal of Physics A: Mathematical and Theoretical*, 46 (50):505202, 2013.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

H. J. Brascamp and E. H. Lieb. Best constants in Young's inequality, its converse, and its generalization to more than three functions. *Advances in Mathematics*, 20(2):151–173, 1976a.

H. J. Brascamp and E. H. Lieb. On extensions of the Brunn-Minkowski and prékopa-leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22(4):366–389, 1976b.

M. Braun and J. McAuliffe. Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335, 2010.

S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

N. Brosse, A. Durmus, É. Moulines, and M. Pereyra. Sampling from a log-concave distribution with compact support with proximal langevin monte carlo. In *Conference on Learning Theory*, pages 319–342, 2017.

N. Brosse, É. Moulines, and A. Durmus. The promises and pitfalls of stochastic gradient langevin dynamics. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8278–8288, 2018.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

X. Cheng and P. Bartlett. Convergence of langevin mcmc in kl-divergence. In *Algorithmic Learning Theory*, pages 186–211. PMLR, 2018.

X. Cheng, N. S. Chatterji, Y. Abbasi-Yadkori, P. L. Bartlett, and M. I. Jordan. Sharp convergence rates for langevin dynamics in the nonconvex setting. *arXiv preprint arXiv:1805.01648*, 2018a.

X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped langevin mcmc: A non-asymptotic analysis. In *Conference on Learning Theory*, pages 300–323. PMLR, 2018b.

X. Cheng, P. L. Bartlett, and M. I. Jordan. Quantitative w_1 convergence of langevin-like stochastic processes with non-convex potential state-dependent noise. *arXiv preprint arXiv:1907.03215*, 2019.

X. Cheng, D. Yin, P. Bartlett, and M. Jordan. Stochastic gradient and langevin processes. In *International Conference on Machine Learning*, pages 1810–1819. PMLR, 2020.

S. Chewi, T. Le Gouic, C. Lu, T. Maunu, and P. Rigollet. Svgd as a kernelized wasserstein gradient flow of the chi-squared divergence. *Advances in Neural Information Processing Systems*, 33: 2098–2109, 2020.

M.-C. Corbineau, D. Kouamé, E. Chouzenoux, J.-Y. Tourneret, and J.-C. Pesquet. Preconditioned p-ula for joint deconvolution-segmentation of ultrasound images. *IEEE Signal Processing Letters*, 26(10):1456–1460, 2019.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

A. Dalalyan. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. In *Conference on Learning Theory*, pages 678–689. PMLR, 2017a.

A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017b.

A. S. Dalalyan and A. Karagulyan. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

A. S. Dalalyan and L. Riou-Durand. On sampling from a log-concave density using kinetic langevin diffusions. *Bernoulli*, 26(3):1956–1988, 2020.

A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and langevin monte-carlo. In *The 22nd Conference on Learning Theory, (COLT)*, pages 1–10, 2009.

A. S. Dalalyan, L. Riou-Durand, and A. Karagulyan. Bounding the error of discretized langevin algorithms for non-strongly log-concave targets. *arXiv preprint arXiv:1906.08530*, 2019.

A. Davis. A differential equation approach to linear combinations of independent chi-squares. *Journal of the American Statistical Association*, 72(357):212–214, 1977.

V. De Bortoli, A. Durmus, M. Pereyra, and A. F. Vidal. Efficient stochastic optimisation by unadjusted langevin monte carlo. *Statistics and Computing*, 31(3):1–18, 2021.

S. S. Du, J. D. Lee, Y. Tian, A. Singh, and B. Poczos. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. In *International Conference on Machine Learning*, pages 1338–1347, 2018.

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

B. Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press, 2012.

M. A. Erdogdu and R. Hosseinzadeh. On the convergence of langevin monte carlo: The interplay between tail growth and smoothness. In *Conference on Learning Theory*, pages 1776–1822. PMLR, 2021.

D. L. Ermak. A computer simulation of charged particles in solution. i. technique and equilibrium properties. *The Journal of Chemical Physics*, 62(10):4189–4196, 1975.

J. Friedman and B. Popescu. Gradient directed regularization. URL `http://www-stat.stanford.edu/~jhf/ftp/pathlite.pdf`. Working paper, 2004.

S. Gabler and C. Wolff. A quick and easy approximation to the distribution of a sum of weighted chi-square variables. *Statistische Hefte*, 28(1):317–325, 1987.

S. B. Gelfand and S. K. Mitter. Recursive stochastic algorithms for global optimization in r^d. *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

J. K. Ghosh, M. Delampady, and T. Samanta. *An introduction to Bayesian analysis: theory and methods*, volume 725. Springer, 2006.

C. R. Givens and R. M. Shortt. A class of wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

N. Gozlan and C. Léonard. Transport inequalities. a survey. *Markov Processes and Related Fields*, 16:635–736, 2010.

U. Grenander. Tutorial in pattern theory. *Report, Division of Applied Mathematics, Brown University*, 1983.

U. Grenander and M. I. Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(4):549–581, 1994.

L. Gross. Logarithmic sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, 2018.

J. Gurland. Distribution of quadratic forms and ratios of quadratic forms. *Ann. Math. Statist.*, 24 (3):416–427, 09 1953.

L. Hodgkinson, R. Salomone, and F. Roosta. Implicit langevin algorithms for sampling from log-concave densities. *Journal of Machine Learning Research*, 22(136):1–30, 2021.

Y.-P. Hsieh, A. Kavis, P. Rolland, and V. Cevher. Mirrored langevin dynamics. In *Advances in Neural Information Processing Systems*, pages 2878–2887, 2018.

D. R. Jensen and H. Solomon. A gaussian approximation to the distribution of a definite quadratic form. *Journal of the American Statistical Association*, 67(340):898–902, 1972.

Z. Ji and M. Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798, 2019.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

A. Karagulyan and A. Dalalyan. Penalized langevin dynamics with vanishing penalty for smooth and log-concave targets. *Advances in Neural Information Processing Systems*, 33, 2020.

P. E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*, volume 23. Springer Science & Business Media, 2013.

V. Lakshmikantham, S. Leela, and A. A. Martynyuk. *Stability analysis of nonlinear systems*. Springer, 1989.

D. Lamberton and G. Pages. Recursive computation of the invariant distribution of a diffusion. *Bernoulli*, 8(3):367–405, 2002.

H. Lee, O. Mangoubi, and N. K. Vishnoi. Online sampling from log-concave distributions. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 1228–1239, 2019.

Y.-A. Ma, N. Chatterji, X. Cheng, N. Flammarion, P. Bartlett, and M. I. Jordan. Is there an analog of nesterov acceleration for mcmc? *arXiv preprint arXiv:1902.00996*, 2019a.

Y.-A. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I. Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42):20881–20885, 2019b.

S. Mandt, M. D. Hoffman, and D. M. Blei. Continuous-time limit of stochastic gradient descent revisited. *NIPS-2015*, 2015.

J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas. A stochastic newton mcmc method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34(3):A1460–A1487, 2012.

E. Mazumdar, A. Pacchiano, Y. Ma, M. Jordan, and P. Bartlett. On approximate thompson sampling with langevin algorithms. In *International Conference on Machine Learning*, pages 6797–6807. PMLR, 2020.

N. Morgan and H. Bourlard. Generalization and parameter estimation in feedforward nets: Some experiments. In *Proceedings of the 2nd International Conference on Neural Information Processing Systems*, pages 630–637, 1989.

W. Mou, Y.-A. Ma, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan. High-order langevin diffusion yields an accelerated mcmc algorithm. *Journal of Machine Learning Research*, 22(42):1–41, 2021.

M. S. Nacson, N. Srebro, and D. Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3051–3059. PMLR, 2019.

R. M. Neal. Bayesian learning via stochastic dynamics. In *Advances in neural information processing systems*, pages 475–482, 1993.

R. M. Neal. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

G. Neu and L. Rosasco. Iterate averaging as regularization for stochastic gradient descent. In *Conference On Learning Theory*, pages 3222–3242. PMLR, 2018.

D. Nguyen, X. Dang, and Y. Chen. Non-convex weakly smooth langevin monte carlo using regularization. *arXiv preprint arXiv:2101.06369*, 2021.

B. Øksendal. *Stochastic differential equations*. Springer, 2003.

A. Orvieto and A. Lucchi. Continuous-time models for stochastic optimization algorithms. *Advances in Neural Information Processing Systems*, 32, 2019.

F. Otto and C. Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

G. Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, 180(3):378–384, 1981.

G. A. Pavliotis. *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer, 2014.

G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

T. Poggio, A. Banburski, and Q. Liao. Theoretical issues in deep networks: Approximation, optimization and generalization. *arXiv preprint arXiv:1908.09375*, 2019.

L. Rademacher and S. Vempala. Dispersion of mass and the complexity of randomized geometric algorithms. *Advances in Mathematics*, 219(3):1037–1069, 2008.

M. Raginsky and J. Bouvrie. Continuous-time stochastic mirror descent on a network: Variance reduction, consensus, convergence. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 6793–6800. IEEE, 2012.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.

J. Ramsay. Parameter flows. Working paper, 2005.

G. Raskutti, M. J. Wainwright, and B. Yu. Early stopping and nonparametric regression: an optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15:335–366, 2014.

H. Robbins and E. Pitman. Application of the method of mixtures to quadratic forms in normal variates. *The annals of mathematical statistics*, pages 552–560, 1949.

C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

G. O. Roberts and R. L. Tweedie. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

P. J. Rossky, J. Doll, and H. Friedman. Brownian dynamics as smart monte carlo simulation. *The Journal of Chemical Physics*, 69(10):4628–4633, 1978.

I. Sato and H. Nakagawa. Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process. In *International Conference on Machine Learning*, pages 982–990, 2014.

S. E. Shreve. *Stochastic calculus for finance II: Continuous-time models*, volume 11. Springer Science & Business Media, 2004.

U. Simsekli, R. Badeau, T. Cemgil, and G. Richard. Stochastic quasi-newton langevin monte carlo. In *International Conference on Machine Learning*, pages 642–651. PMLR, 2016.

R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

H. Solomon and M. A. Stephens. Distribution of a sum of weighted chi-square variables. *Journal of the American Statistical Association*, 72(360a):881–885, 1977.

D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

O. N. Strand. Theory and methods related to the singular-function expansion and landweber's iteration for integral equations of the first kind. *SIAM Journal on Numerical Analysis*, 11(4): 798–825, 1974.

A. Suggala, A. Prasad, and P. K. Ravikumar. Connecting optimization and regularization paths. In *Advances in Neural Information Processing Systems*, pages 10608–10619, 2018.

A. Takatsu. Wasserstein geometry of gaussian measures. *Osaka Journal of Mathematics*, 48(4): 1005–1026, 2011.

M. Talagrand. Transportation cost for gaussian and other product measures. *Geometric & Functional Analysis GAFA*, 6(3):587–600, 1996.

D. Talay and L. Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic analysis and applications*, 8(4):483–509, 1990.

Y. W. Teh, A. H. Thiery, and S. J. Vollmer. Consistency and fluctuations for stochastic gradient langevin dynamics. *The Journal of Machine Learning Research*, 17(1):193–225, 2016.

T. Vaskevicius, V. Kanade, and P. Rebeschini. Implicit regularization for optimal sparse recovery. In *Advances in Neural Information Processing Systems*, pages 2968–2979, 2019.

P. Vatiwutipong and N. Phewchean. Alternative way to derive the distribution of the multivariate ornstein–uhlenbeck process. *Advances in Difference Equations*, 2019(1):1–7, 2019.

R. Vershynin. High-dimensional probability, volume 47 of cambridge series in statistical and probabilistic mathematics, 2018.

C. Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Y. Wang and S. Wu. Asymptotic analysis via stochastic differential equations of gradient descent algorithms in statistical and computational paradigms. *J. Mach. Learn. Res.*, 21:199–1, 2020.

Y. Wei, F. Yang, and M. J. Wainwright. Early stopping for kernel boosting algorithms: a general analysis with localized complexities. In *Advances in Neural Information Processing Systems*, 2017.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

A. Wibisono. Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pages 2093–3027. PMLR, 2018.

X. Wu, E. Dobriban, T. Ren, S. Wu, Z. Li, S. Gunasekar, R. Ward, and Q. Liu. Implicit regularization and convergence for weight normalization. *Advances in Neural Information Processing Systems*, 33, 2020.

P. Xu, T. Wang, and Q. Gu. Continuous and discrete-time accelerated stochastic mirror descent for strongly convex functions. In *International Conference on Machine Learning*, pages 5488–5497, 2018.

Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

D. Zou, P. Xu, and Q. Gu. Faster convergence of stochastic gradient langevin dynamics for non-log-concave sampling. *arXiv preprint arXiv:2010.09597*, 2020.