A Non-Asymptotic Analysis of Stochastic Gradient Langevin Dynamics

Tyler Farghly
tyler.farghly@stats.ox.ac.uk

November 2020

Stochastic Gradient Langevin Dynamics is a popular variant of Stochastic Gradient Descent where Gaussian noise is added at each iteration. This simple addition allows it to escape local minima and obtain impressive results for non-convex objectives. In these notes, we review ideas from the paper "Non-Convex Learning via Stochastic Gradient Langevin Dynamics: A Nonasymptotic Analysis" [6]. In this paper, Raginsky et al. apply powerful tools from probability theory to obtain generalization bounds in the setting of non-convex learning.

1 Introduction

We begin by establishing some notation and introducing some basic concepts.

1.1 The Setting

Consider a standard prediction problem with the set of actions $\mathbb{R}^d$ and examples $\mathcal{Z}$ (for example it could be that $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ and $\mathcal{Y}$ are the sets of possible features and labels). Let $f : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}_+$ be a loss function, then given a distribution on the space $\mathcal{Z}$ we define the *risk*, $r : \mathbb{R}^d \to \mathbb{R}_+$ where $r(w) = \mathbb{E}_Z f(w, Z)$. The goal is to minimize r which is usually intractable and so we instead minimize the *empirical risk*, $R : \mathbb{R}^d \to \mathbb{R}_+$. Given a collection of n examples $\mathbf{z} = (z_1, ..., z_n) \in \mathcal{Z}^n$, this is defined by

$$R(w) = \frac{1}{n} \sum_{i=1}^{n} f(w, z_i).$$

In the case that for each $z \in \mathcal{Z}$, $f(\cdot, z) \in C^1(\mathbb{R}^d)$, a popular approach to minimizing R is by using *gradient descent methods*. Since n is usually large in the non-convex setting, we approximate the gradient ∇R with a stochastic approximation resulting in the update:

$$W_{k+1} = W_k - \eta g_k(W_k),$$

where g_k forms a sequence of random variables, unbiased estimators for the function ∇R, and $\eta > 0$ is the learning rate. In these notes we consider only the *mini-batch* approximation, given by

$$g_k(w) = \frac{1}{n_m} \sum_{i \in U_k} f(w, z_i).$$

where for each k, U_k is a uniform sample from the set $\{U : U \subset [n], |U| = n_m\}$ and $n_m > 0$ is the mini-batch size.

1.2 Stochastic Gradient Langevin Dynamics

Stochastic Gradient Langevin Dynamics (SGLD) is obtained from the above algorithm by simply adding appropriately scaled isotropic Gaussian noise to the update. The relevant update reads

$$W_{k+1} = W_k - \eta g(W_k, U_k) + \sqrt{2\beta^{-1}}\xi_k,$$

where for each k, ξ_k is independently and identically distributed according to $\xi_k \sim N(0, I_d)$ and is independent from $(g_k)_{k\geq 0}$. Heuristically, it can be argued that this modification prevents the process from becoming stuck around local minima and so is a natural addition in the non-convex case. However, there is the antithesis that once W_k is near a minimizer of R the noise term could prevent it from properly converging.

A more refined argument is given in [3, 9] where it is pointed out that if we take η to its infinitesimal limit the process W_k can be approximated by $W(k\eta)$, where $W(t)$ is the solution to the stochastic differential equation (SDE)

$$dW(t) = -\nabla R(W(t))dt + \sqrt{2\beta^{-1}}dB(t), \quad W(0) = W_0. \tag{1}$$

Here $B(t)$ is a standard d-dimensional Wiener process and is independent of W_0. Equation (1) is the (overdamped) *Langevin equation* or *Smoluchowski SDE* [5] and has been studied extensively. A central property is that if R is suitably regular we can guarantee that for a suitable measurable function $\varphi : \mathbb{R}^d \to \mathbb{R}$,

$$\mathbb{E}\varphi(W(t)) \to \hat{\pi}(\varphi), \text{ as } t \to \infty, \tag{2}$$

where $\hat{\pi}$ is the *Gibbs measure* for the potential βR:

$$\hat{\pi}(dw) = \frac{1}{\Lambda} e^{-\beta R(w)} dw, \quad \Lambda = \int_{\mathbb{R}^d} e^{-\beta R(w)} dw. \tag{3}$$

Here and throughout the report, we use the notation $\hat{\pi}(\varphi)$ to denote the integral of φ with respect to the measure $\hat{\pi}$. This has great importance in the optimization setting since $\hat{\pi}$ is concentrated around w where $R(w)$ is small. Furthermore, if R has a single minimizer w^* it can be shown that as $\beta \to \infty$, $\hat{\pi}$ converges weakly to the Dirac measure δ_{w^*}. This motivates the *simulated annealing* algorithm which gradually decreases β^{-1} to 0. This method has shown great promise in many settings, but has proven difficult in practice due to it being overly sensitive to the rate at which β^{-1} converges.

1.3 The Main Result

We now turn to the paper that forms the subject of this report. The core aim is to understand the statistical properties of SGLD, in particular how effectively it can minimize the risk. The novelty in this work lies in the fact that the analysis is *non-asymptotic* and focuses on finite-time guarantees. If the limiting properties of the Langevin equation given in the previous section are to be used they must first be refined into non-asymptotic results which naturally will require the adoption of additional analytic tools.

Another important aspect of this analysis is that it doesn't restrict the loss function to the convex case. With SGLD finding popularity primarily in non-convex learning problems it is fitting to avoid such a constraint.

The main result of the paper[6] is, under the suitable assumptions and for suitably small $\varepsilon > 0$ we have

$$\mathbb{E}r(W_k) - r^* \leq \mathcal{O}\bigg(c_l(\beta, d) \cdot \Big(\varepsilon + n^{-1/4}\log(\varepsilon) + \beta n^{-1}\Big) + \frac{d\log(\beta)}{\beta}\bigg), \tag{4}$$

where $k = \Omega\big(c(\beta, d)\varepsilon^{-4}\log\big(1/\varepsilon\big)\big)$ and $\eta^{1/4} \leq \mathcal{O}\big(\varepsilon \log\big(1/\varepsilon\big)\big)$. The constant $c_l(\beta, d)$ is the *logarithmic Sobolev constant* and is bounded by

$$c_l(\beta, d) \leq \mathcal{O}\big(\beta + d\big)^2 \exp(\mathcal{O}\big(\beta + d\big)\big).$$

To obtain this result they show that W_k and $W(t)$ are *close* in some precises sense and then perform most of the analysis using the Langevin equation. In the interest of brevity we will focus only on the excess risk for the Langevin dynamics. This is extended to the full result following the addition of a single term (the $n^{-1/4}\log(\varepsilon)$ term in the equation above).

2 Functional Inequalities and Exponential Ergodicity

In this section we will discuss logarithmic Sobolev inequalities and some of their interesting properties. Speaking generally, functional inequalities of this sort are designed to provide an abstract link between analytic features of evolving systems and probabilistic convergence to equilibrium. We will consider only the case of the Langevin dynamics given in (1) as a more broadly applicable presentation would require the use of diffusion semigroups which are outside of the scope of these notes (see appendix A).

2.1 Logarithmic Sobolev inequalities

For a probability measure μ on $\mathbb{R}^d$ and a suitably integrable function $\varphi : \mathbb{R}^d \to \mathbb{R}_+$ define the *entropy*

$$\mathrm{Ent}_\mu(\varphi) = \mu\big(\varphi \log \varphi\big) - \mu(\varphi) \log \mu(\varphi).$$

This can produce many familiar quantities. For example suppose φ is a probability density with respect to the Lebesgue measure, $\nu(dx) = \varphi(x)dx$, then the entropy gives the negative *differential entropy*,

$$\mathrm{Ent}_\lambda(\varphi) = -h(\nu) = \int_{\mathbb{R}^d} \varphi(x) \log \varphi(x) dx,$$

where λ is the Lebesgue measure on $\mathbb{R}^d$. Also, if $\varphi = \mathrm{d}\nu/\mathrm{d}\mu$ is the Radon-Nikodym derivative then we recover the *relative entropy*

$$Ent_\mu(\varphi) = D(\nu \| \mu) = \int_{\mathbb{R}^d} \frac{\mathrm{d}\nu}{\mathrm{d}\mu} \log \frac{\mathrm{d}\nu}{\mathrm{d}\mu} d\mu.$$

We say that $\hat{\pi}$ satisfies a logarithmic Sobolev inequality with constant $c_l > 0$ if for any $\varphi \in C^1(\mathbb{R}^d)$

$$\mathrm{Ent}_{\hat{\pi}}(\varphi^2) \leq 2c_l \beta^{-1} \hat{\pi}(\|\nabla \varphi\|^2).$$

The purpose of this inequality is not clear at first glance. Speaking abstractly, it can be seen from the definition of $\hat{\pi}$ that it reduces to a property of the potential R and the parameter β. Once we show that such an inequality is related to the probabilistic convergence of $W(t)$ through the parameter c we will have a direct, albeit complicated, relationship between the geometry of R, the amount of noise applied and the rate of convergence.

2.2 Exponential Ergodicity and Other Properties

An important result to follow from such an inequality is exponential convergence in entropy. For all $t \geq 0$ define the probability measure ν_t such that $W(t) \sim \nu_t$.

Theorem 2.1 (Exponential ergodicity in relative entropy [1])**.** *Suppose $\hat{\pi}$ satisfies a logarithmic Sobolev inequality with constant $c > 0$. Then if $D(\nu_0 \| \hat{\pi}) < \infty$ we have exponential convergence in relative entropy:*

$$D(\nu_t \| \hat{\pi}) \leq e^{-2t/c} D(\nu_0 \| \hat{\pi}).$$

This theorem is a non-asymptotic analogue of the convergence result stated in (2). In Proposition 3.5 we will use this result to obtain non-asymptotic bounds in expectations of functions. It can be also be shown that this gives convergence in second-moment: $\nu_t(\|w\|^2) \to \hat{\pi}(\|w\|^2)$ as $t \to \infty$ (see Theorem 7.12 in [8]).

Another interesting property of this inequality is its relationship to the problem of optimal transportation. Let μ and ν be probability measures on $\mathbb{R}^d$ and define the *coupling* of μ and ν, $\Gamma(\mu, \nu)$ as the set of probability measures γ on $\mathbb{R}^d \times \mathbb{R}^d$ such that if $(X, Y) \sim \gamma$ it has marginal distributions $X \sim \mu$, $Y \sim \nu$. Define the 2-*Wasserstein distance* between μ and ν by

$$\mathcal{W}_2(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int \|x - y\|^2 \gamma(dx, dy) \right)^{\frac{1}{2}}.$$

The problem of optimal transportation is not especially relevant to the subject of these notes. We are only interested in this object because it has some pleasant computational properties. For example, unlike the relative entropy this is a metric and so it admits nice properties such as the triangle inequality. We form a relationship between these two objects using logarithmic Sobolev inequalities.

Theorem 2.2 (Otto-Villani Theorem [1])**.** *Suppose $\hat{\pi}$ satisfies a logarithmic Sobolev inequality with constant $c_l > 0$. Then for any probability measure μ on $\mathbb{R}^d$,*

$$\mathcal{W}_2(\hat{\pi}, \mu)^2 \leq 2c_l D(\mu \| \hat{\pi}).$$

There are many more interesting properties that follow from this inequality but the two given above are the only ones that will be relevant in the following analysis. Another application of this inequality can be found in the study of concentration which is covered in [7].

3 The Diffusion Approximation of the Gibbs Sampler

We now return to the primary subject of these notes. In this section we treat the training data $\mathbf{z}$ as a deterministic element of $\mathcal{Z}^n$ which we extend to the random case in the sequel.

3.1 Optimization Error of the Langevin Dynamics

Let's start by assuming the existence of a logarithmic Sobolev inequality – later we will give sufficient conditions for it to hold:

(B.1) Suppose $\hat{\pi}$ satisfies a logarithmic Sobolev inequality with constant c_l.

As we have seen in the previous section, under the assumption of (B.1) we have some powerful and relevant results – most notably exponential ergodicity in relative entropy. The relevance of this is made clear by the following decomposition:

Lemma 3.1. *Suppose μ is a probability measure on $\mathbb{R}^d$ with finite differential entropy such that $\mu(R) < \infty$. Then we have*

$$D(\mu\|\hat{\pi}) = -h(\mu) + \beta\mu(R) + \log(\Lambda).$$

This follows immediately from the definition of relative entropy. We can apply this with $\mu = \nu_t$ and $\mu = \hat{\pi}$ to obtain

$$\nu_t(R) - \hat{\pi}(R) = \beta^{-1}\big(D(\nu_t\|\hat{\pi}) + h(\nu_t) - h(\hat{\pi})\big). \tag{5}$$

Here we have used the fact that $D(\hat{\pi}\|\hat{\pi}) = 0$.

A simple way of bounding the differential entropy follows from the *maximum entropy principle*. This states that given a probability distribution on $\mathbb{R}^d$ has finite total variance σ^2, it must have differential entropy bounded by that of the Gaussian random variable $\xi \sim N(0, \sigma^2 I_d)$ given by

$$h(\xi) = \frac{d}{2}\log(2\pi e\sigma^2). \tag{6}$$

In the next proposition we will make use of this bound. This also motivates the next assumption:

(B.2) Suppose there exists $B_c > 0$ such that for any $\mathbf{z} \in \mathcal{Z}^N$, $t \geq 0$ it holds that $\mathbb{E}(\|W(t)\|^2) \leq B_c$.

In addition, we will need to make some assumptions about the smoothness of R:

(B.3) Suppose there exists $M, B > 0$ such that for all $z \in \mathcal{Z}$, $f(\cdot, z)$ is M-smooth and $\|\nabla f(0, z)\| \leq B$.

Recall that a function $\varphi \in C^1(\mathbb{R}^d)$ is M-smooth if for any $w_1, w_2 \in \mathbb{R}^d$

$$\|\nabla\varphi(w_1) - \nabla\varphi(w_2)\| \leq M\|w_1 - w_2\|.$$

An immediate consequence of this assumption is a bound on the gradient that is uniform over $z \in \mathcal{Z}$:

$$\|\nabla f(w, z)\| \leq M\|w\| + B. \tag{7}$$

It follows from this assumption that both R and r are also M-smooth and have gradients bounded by B at the origin.

Lemma 3.2. *Suppose assumption (B.3) holds, then for any $w \in \mathbb{R}^d$ and $w^* \in \operatorname{argmin}_{w\in\mathbb{R}^d} R(w)$ it holds that*

$$|R(w) - R(w^*)| \leq \frac{M}{2}\|w - w^*\|^2.$$

Proof. By the fundamental theorem of calculus we obtain

$$R(w) - R(w^*) = \int_0^1 \langle w - w^*, \nabla R(tw^* + (1-t)w)\rangle dt.$$

R is M-smooth and $\nabla R(w^*) = 0$ so we can bound the gradient by $\|\nabla R(w)\| \leq M\|w - w^*\|$. After applying the Cauchy-Schwarz inequality we use this bound to deduce

$$|R(w) - R(w^*)| \leq \int_0^1 \|w - w^*\| \|\nabla R(tw^* + (1-t)w)\| dt \leq \int_0^1 Mt\|w - w^*\|^2 dt.$$

□

Following from a similar approach to the result given in (5), we now deduce bounds in expectation for the optimization error.

Proposition 3.3. *Suppose (B.1), (B.2) and (B.3) hold and the distribution of W_0 has the property $D(\mu_0\|\hat{\pi}) < \infty$. Then we have the bound*

$$\mathbb{E}R\big(W(t)\big) - R^* = \beta^{-1}e^{-2t/c_l}D(\mu_0\|\hat{\pi}) + \frac{d}{2}\beta^{-1}\log(MB_c\beta e),$$

where $R^* = \min_{w\in\mathbb{R}^d} R(w)$.

Note that the first term decays exponentially fast in t while the second term is fixed. This result suggests that the stopping time should grow linearly with c_l.

Proof. Applying Lemma 3.1 to the measure ν_t and subtracting R^* yields

$$\mathbb{E}R\big(W(t)\big) - R^* = \beta^{-1}\big(D(\nu_t\|\hat{\pi}) + h(\nu_t) - \log(\Lambda) - \beta R^*\big).$$

We start by bounding the last two terms which can be written as

$$\log(\Lambda) + \beta R^* = \log\left(\int_{\mathbb{R}^d} e^{-\beta(R(w)-R^*)}dw\right).$$

By assumption Lemma 3.2 the exponent is bounded using $R(w) - R^* \leq M/2\|w - w^*\|^2$ which holds for any $w^* \in \operatorname{argmin}_{w\in\mathbb{R}^d} R(w)$. With this, we compute the Gaussian integral yielding

$$\log(\Lambda) + \beta R^* \geq \frac{d}{2}\log\left(\frac{2\pi}{\beta M}\right). \tag{8}$$

Next, the maximum entropy principle is used bound the second term. Certainly (B.2) provides B_c as a bound on the variance of ν_t and hence

$$h(\nu_t) \leq \frac{d}{2}\log(2\pi e B_c). \tag{9}$$

To conclude, we use Theorem 2.1 to bound the first term. □

The constant B_c describes how confined the process $W(t)$ is to the origin. Certainly if R was convex with large gradients when moving away from the origin, B_c would be small. However, such conditions would also result in a large smoothness parameter M. The previous result points to a tradeoff between the benefits of having a relatively flat objective (resulting in a low value of M) and those of having a sharp confining objective (a low value for B_c).

3.2 Obtaining Error in Population Risk

In the previous section we exploited the fact that R appears in the density of $\hat{\pi}$ to form a relationship between the quantity $\mathbb{E}R(W(t))$ and the divergence $D(\nu_t\|\hat{\pi})$. Thus to analyse the quantity $\mathbb{E}r(W(t))$ a different approach is required.

However the principle exhibited in (5), that $\nu_t(R) - \hat{\pi}(R)$ should be small since ν_t and $\hat{\pi}$ are *close* in some sense and R is *nice* in some sense, should hold once R is replaced with any suitably *nice* function.

Lemma 3.4. *(Wasserstein continuity for quadratic bounded functions). Suppose μ and ν are probability measures on $\mathbb{R}^d$ with finite second moments and $\varphi \in C^1(\mathbb{R}^d)$ such that for some $c_1, c_2 > 0$, $\|\nabla\varphi(w)\| \leq c_1\|w\| + c_2$ for all $w \in \mathbb{R}^d$. Then there is the bound*

$$\left|\int_{\mathbb{R}^d} \varphi(w)\mu(dw) - \int_{\mathbb{R}^d} \varphi(w)\nu(dw)\right| \leq (c_1\sigma + c_2)\mathcal{W}_2(\mu,\nu), \tag{10}$$

where $\sigma^2 = \int_{\mathbb{R}^d} \|w\|^2\mu(dw) \vee \int_{\mathbb{R}^d} \|w\|^2\nu(dw)$.

Proof. By the fundamental theorem of calculus, for any $w, v \in \mathbb{R}^d$ we have

$$\varphi(w) - \varphi(v) = \int_0^1 \langle w - v, \nabla\varphi(tw + (1-t)v)\rangle dt.$$

Applying the Cauchy-Schwarz inequality along with the quadratic bounds yields

$$|\varphi(w) - \varphi(v)| \leq \|w - v\|\left(\frac{c_1}{2}(\|w\| + \|v\|) + c_2\right). \tag{11}$$

Recall the notion of a coupling from section 2.1. Taking any $\gamma \in \Gamma(\mu,\nu)$, the right-hand side of 10 can be written as a single integral yielding

$$\begin{aligned}\left|\int_{\mathbb{R}^d} \varphi(w)\mu(dw) - \int_{\mathbb{R}^d} \varphi(w)\nu(dw)\right|^2 &\leq \left[\int_{\mathbb{R}^d} |\varphi(w) - \varphi(v)|\gamma(dw,dv)\right]^2 \\ &\leq \int_{\mathbb{R}^d} \|w - v\|^2\gamma(dw,dv)\int_{\mathbb{R}^d}\left(\frac{c_1}{2}(\|w\| + \|v\|) + c_2\right)^2\gamma(dw,dv),\end{aligned}$$

where the last line follows from equation (11) and the Cauchy-Schwarz inequality. The second integral is bounded using the triangle inequality in $L^2(\gamma)$:

$$\begin{aligned}\left[\int_{\mathbb{R}^d}\left(\frac{c_1}{2}(\|w\| + \|v\|) + c_2\right)^2\gamma(dw,dv)\right]^{1/2} &\leq \frac{c_1}{2}\left(\sqrt{\int_{\mathbb{R}^d}\|w\|^2\mu(dw)} + \sqrt{\int_{\mathbb{R}^d}\|w\|^2\nu(dw)}\right) + c_2 \\ &\leq c_1\sigma + c_2.\end{aligned}$$

Finally, taking an infimum of the first term over all $\gamma \in \Gamma(\mu,\nu)$ gives $\mathcal{W}_2^2(\mu,\nu)$. □

With this we can now bound the expected difference in population risks of $W(t)$ and the Gibbs algorithm which will be useful for computing the excess risk. Certainly, after we obtain the excess risk of the Gibbs algorithm we can add this term to obtain the excess risk of $W(t)$.

Proposition 3.5. *Suppose (B.1), (B.2) and (B.3) hold and $D(\mu_0\|\hat{\pi}) < \infty$. Then there is the bound*

$$|\nu_t(r) - \hat{\pi}(r)| \leq \left(M\sqrt{B_c} + B\right)\sqrt{2c_l D(\mu_0\|\hat{\pi})}\, e^{-t/c_l}.$$

Proof. Since $\hat{\pi}$ satisfies a logarithmic Sobolev inequality, Theorems 2.1 and 2.2 combine to give the bound

$$\mathcal{W}_2(\nu_t,\hat{\pi}) \leq \sqrt{2c_l D(\mu_0\|\hat{\pi})}\, e^{-t/c_l}.$$

The result then follows from Lemma 3.4 with $\varphi = R$, $\mu = \nu_t$, $\nu = \hat{\pi}$. Indeed, by assumption R satisfies the quadratic bounds with $c_1 = M$ and $c_2 = B$. To obtain a bound for σ^2 we recall from section 2.1 that $\nu_t(\|w\|^2) \to \hat{\pi}(\|w\|^2)$ as $t \to \infty$. Therefore it must hold that $\hat{\pi}(\|w\|^2) \leq B_c$ and so $\sigma^2 \leq B_c$. □

Showing the asymptotic result $\nu_t(r) \to \hat{\pi}(r)$ as $t \to \infty$ requires far less work. It doesn't require the satisfaction of any functional inequality or bounded variance and only depends on R being sufficiently regular [5]. Indeed, the cost of the non-asymptotic analysis is the requirement for more aggressive assumptions and more complex tools.

4 Generalization Properties of the Gibbs Algorithm

In the previous section we considered a fixed set of training examples $\mathbf{z}$. If we are to study the generalization properties of this algorithm we must understand the stability of it under changes in the set of training examples. To make the dependence on the examples explicit we use the notation $\hat{\pi}_{\mathbf{z}}$ to denote the Gibbs measure with examples $\mathbf{z}$ and $\Lambda_{\mathbf{z}}$ to denote the corresponding normalization constant.

4.1 Uniform Stability

We start by studying the stability of the Gibbs algorithm under changes to single coordinates of the example set.

Proposition 4.1. *Suppose (B.1), (B.2) and (B.3) hold and $\bar{\mathbf{z}} \in \mathcal{Z}^n$ differs from $\mathbf{z}$ in a single coordinate. Then the associated Gibbs measures for $\mathbf{z}$ and $\bar{\mathbf{z}}$ have the property*

$$D(\hat{\pi}_{\bar{\mathbf{z}}} \| \hat{\pi}_{\mathbf{z}}) \leq \frac{2c_l\beta^2}{n^2}\big(M^2 B_c + B^2\big)$$

Proof. Suppose $\mathbf{z}$ and $\bar{\mathbf{z}}$ differ in only the coordinate of index i, for which they take the values z_i and $\bar{z}_i$ respectively. Consider the Radon-Nikodym derivative derivative of $\pi_{\mathbf{z}}$ with respect to $\pi_{\bar{\mathbf{z}}}$:

$$\frac{\mathrm{d}\hat{\pi}_{\bar{\mathbf{z}}}}{\mathrm{d}\hat{\pi}_{\mathbf{z}}}(w) = \frac{\Lambda_{\mathbf{z}}}{\Lambda_{\bar{\mathbf{z}}}} \exp\left(\frac{\beta}{n}(f(w, z_i) - f(w, \bar{z}_i))\right).$$

Since by assumption $\hat{\pi}_{\mathbf{z}}$ satisfies a logarithmic Sobolev inequality it follows that

$$D(\hat{\pi}_{\bar{\mathbf{z}}} \| \hat{\pi}_{\mathbf{z}}) = Ent_{\hat{\pi}_{\mathbf{z}}}\left(\frac{\mathrm{d}\hat{\pi}_{\bar{\mathbf{z}}}}{\mathrm{d}\hat{\pi}_{\mathbf{z}}}\right) \leq 2c_l \int_{\mathbb{R}^d} \left\| \nabla \sqrt{\frac{\mathrm{d}\hat{\pi}_{\bar{\mathbf{z}}}}{\mathrm{d}\hat{\pi}_{\mathbf{z}}}}(w) \right\|^2 \hat{\pi}_{\mathbf{z}}(dw). \tag{12}$$

To this end, we compute the gradient

$$\nabla \sqrt{\frac{\mathrm{d}\hat{\pi}_{\bar{\mathbf{z}}}}{\mathrm{d}\hat{\pi}_{\mathbf{z}}}}(w) = \frac{\beta}{2n}\big(\nabla f(w, z_i) - \nabla f(w, \bar{z}_i)\big) \sqrt{\frac{\mathrm{d}\hat{\pi}_{\bar{\mathbf{z}}}}{\mathrm{d}\hat{\pi}_{\mathbf{z}}}(w)}.$$

Substituting this into (12) yields

$$\begin{aligned} D(\hat{\pi}_{\bar{\mathbf{z}}} \| \hat{\pi}_{\mathbf{z}}) &\leq \frac{c_l\beta^2}{2n^2} \int_{\mathbb{R}^d} \|\nabla f(w, z_i) - \nabla f(w, \bar{z}_i)\|^2 \hat{\pi}_{\bar{\mathbf{z}}}(dw) \\ &\leq \frac{2c_l\beta^2}{n^2}\Big(M^2\, \hat{\pi}_{\bar{\mathbf{z}}}\big(\|w\|^2\big) + B^2\Big), \end{aligned}$$

where the second line follows from (7), a consequence of (B.3). Using assumption (B.2) we can further bound the second-moment term with $\hat{\pi}_{\bar{\mathbf{z}}}\big(\|w\|^2\big) \leq B_c$. □

Using Lemma 3.4 and the Otto-Villani theorem we readily deduce uniform stability for the Gibbs algorithm.

Corollary 4.2. *Suppose (B.1), (B.2) and (B.3) hold and $\bar{\mathbf{z}} \in \mathcal{Z}^n$ differs from $\mathbf{z}$ in a single coordinate. For any $z \in \mathcal{Z}$ we have the stability bound*

$$\big|\hat{\pi}_{\mathbf{z}}(f(\cdot, z)) - \hat{\pi}_{\bar{\mathbf{z}}}(f(\cdot, z))\big| \leq \frac{4c_l\beta}{n}\big(M^2 B_c + B^2\big).$$

Certainly, it is expected that sensitivity to a change in the training examples should reduce as the number of training examples grows. In the next section, we will show that this result allows us to bound the expected difference in the risk and empirical risk for the Gibbs algorithm.

4.2 Generalization Error Bounds

To obtain bounds in generalization error it is vital that we treat the training examples $\mathbf{Z} = (Z_1, ..., Z_N)$ as a random variable (as highlighted by its capitalisation). Suppose each coordinate is independently and identically distributed according to the example distribution P on the set $\mathcal{Z}$. Note that now the empirical risk R and the measure ν_t are both random.

Proposition 4.3. *Suppose (B.1), (B.2) and (B.3) hold with identical constants for any $\mathbf{z} \in \mathcal{Z}^n$. Then there is the bound,*

$$\mathbb{E}_{\mathbf{Z}}\big(\hat{\pi}_{\mathbf{Z}}(r) - \hat{\pi}_{\mathbf{Z}}(R)\big) \leq \frac{4c_l\beta}{n}\big(M^2B_c + B^2\big).$$

Proof. Let $\bar{Z} \sim P$ be a random variable that is independent of $\mathbf{Z}$, then the generalization error reads

$$\mathbb{E}_{\mathbf{Z}}\big(\hat{\pi}_{\mathbf{Z}}(r) - \hat{\pi}_{\mathbf{Z}}(R)\big) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{Z}}\mathbb{E}_{\bar{Z}}\hat{\pi}_{\mathbf{Z}}\big(f(\cdot, \bar{Z}) - f(\cdot, Z_i)\big).$$

Since $(Z_1, ..., Z_n, \bar{Z})$ are independently and identically distributed according to P and we are taking the expectation over all of these, the expression above would be identical if $\bar{Z}$ was swapped for any one of the coordinates of $\mathbf{Z}$. To this end, let $\bar{\mathbf{Z}}^i = (\bar{Z}_1^i, ..., \bar{Z}_N^i)$ with $\bar{Z}_j^i = Z_j$ for $j \neq i$ and $\bar{Z}_i^i = \bar{Z}$. From this argument it follows that $\mathbb{E}_{\mathbf{Z}}\mathbb{E}_{\bar{Z}}\hat{\pi}_{\mathbf{Z}}f(\cdot, Z_i) = \mathbb{E}_{\bar{\mathbf{Z}}^i}\mathbb{E}_{Z_i}\hat{\pi}_{\bar{\mathbf{Z}}}f(\cdot, \bar{Z})$ – effectively we have swapped the coordinate of index i with $\bar{Z}$. Since $\mathbf{Z}$ and $\bar{\mathbf{Z}}$ differ at a single coordinate, Corollary 4.2 yields the estimate

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}}\mathbb{E}_{\bar{Z}}\hat{\pi}_{\mathbf{Z}}\big(f(\cdot, \bar{Z}) - f(\cdot, Z_i)\big) &= \mathbb{E}_{\mathbf{Z}}\mathbb{E}_{\bar{Z}}\Big[\hat{\pi}_{\mathbf{Z}}\big(f(\cdot, \bar{Z})\big) - \hat{\pi}_{\bar{\mathbf{Z}}^i}\big(f(\cdot, \bar{Z})\big)\Big] \\ &\leq \frac{4c_l\beta}{n}\big(M^2B_c + B^2\big).\end{aligned}$$

□

Note that the previous result depends on the Gibbs algorithm only by its uniform stability bound and can be applied to any algorithm with this property.

5 Completing the Proof

5.1 Excess Risk of the Langevin Dynamics

Computing the excess risk is now a task of combining these results. Continuing to treat the training examples as a random variable we note that the expected population risk can be written as $\mathbb{E}r\big(W(t)\big) = \mathbb{E}_{\mathbf{Z}}\nu_t(r)$. In fact, we can decompose the excess risk to give:

$$\mathbb{E}r\big(W(t)\big) - r^* = \mathbb{E}_{\mathbf{Z}}\big(\nu_t(r) - \hat{\pi}_{\mathbf{Z}}(r)\big) + \mathbb{E}_{\mathbf{Z}}\big(\hat{\pi}_{\mathbf{Z}}(r) - \hat{\pi}_{\mathbf{Z}}(R)\big) + \mathbb{E}_{\mathbf{Z}}\big(\hat{\pi}_{\mathbf{Z}}(R) - R^*\big) + \mathbb{E}_{\mathbf{Z}}\big(R^* - r^*\big). \tag{13}$$

We have already seen the first term in Proposition 3.5 when the training examples were considered to be deterministic. To extend this result to the non-deterministic case we require that the assumptions hold with the same constants for any choice of training examples $\mathbf{z} \in \mathcal{Z}^n$. In addition, if we assume that there exists some $\gamma > 0$ such that $D(\mu_0\|\hat{\pi}_{\mathbf{z}}) < \gamma$ then the proposition can be readily extended to the uniform bound

$$\mathbb{E}_{\mathbf{Z}}\big(\nu_t(r) - \hat{\pi}_{\mathbf{Z}}(r)\big) \leq \big(M\sqrt{B_c} + B\big)\sqrt{2c_l\gamma}\, e^{-t/c_l}. \tag{14}$$

Once the other terms are handled we obtain the following bound in excess risk.

Proposition 5.1. *Suppose assumptions (B.1), (B.2) and (B.3) hold with identical constants for any $\mathbf{z} \in \mathcal{Z}^n$ and $D(\mu_0\|\hat{\pi}_{\mathbf{z}}) < \gamma$. Then,*

$$\mathbb{E}r\big(W(t)\big) - r^* \leq \sqrt{2\sigma^2c_l\gamma}\, e^{-t/c_l} + \frac{4c_l\beta}{n}\sigma^2 + \frac{d}{2}\beta^{-1}\log(MB_c\beta e)$$

where $\sigma = M\sqrt{B_c} + B$.

To obtain a result similar to the one given in 4 we use a free parameter $\varepsilon \in (0,1)$ and require that $t = c_l \log(1/\varepsilon)$. A formulation of this sort is useful in practice for performing early stopping.

Proof. The third term of (13) is similar to the optimization error considered in Proposition 3.3 and can be bounded similarly. Indeed, for any $\mathbf{z} \in \mathcal{Z}^n$ Lemma 3.1 and $D(\hat{\pi}_{\mathbf{z}} \| \hat{\pi}_{\mathbf{z}}) = 0$ give

$$\hat{\pi}_{\mathbf{z}}(R) - R^* = \beta^{-1}\big(h(\hat{\pi}_{\mathbf{z}}) - \log(\Lambda_{\mathbf{z}}) - \beta R^*\big).$$

Using the same argument as in Proposition 3.5 we obtain $\hat{\pi}_{\mathbf{z}}(\|w\|^2) \leq B_c$ and so we can readily apply the same entropy bounds as in the proof of Proposition 3.3. From this we obtain the estimate

$$\hat{\pi}_{\mathbf{z}}(R) - R^* \leq \frac{d}{2}\beta^{-1}\log(MB_c\beta e).$$

For the fourth term, note that $\mathbb{E}_{\mathbf{z}}(R(w)) = r(w)$ and so for any $w^* \in \operatorname{argmin}_{w\in\mathbb{R}^d} R(w)$ we can write

$$\mathbb{E}_{\mathbf{Z}}\big(R^* - r^*\big) = \mathbb{E}_{\mathbf{Z}}\Big(\min_{w\in\mathbb{R}^d} R(w) - R(w^*)\Big) \leq 0$$

The second and third terms are bounded in equation (14) and Proposition 4.3 respectively. Combining these as in (13) gives the final result. □

One issue with this result is it depends greatly on the quantity c_l which is not easy to interpret. Also, it almost certainly depends on the constants β, M, B_c, d and B without explicitly expressing how. Further, the assumptions made are somewhat technical and so the practical cases in which this result applies is not yet clear. In the next section we will discuss the assumptions given in the original paper and state a bound for the logarithmic Sobolev constant.

5.2 Assumptions and the Logarithmic Sobolev Constant

Showing that a measure of the form $\hat{\pi}$ satisfies a logarithmic Sobolev inequality is not at all easy. Providing simple criterion for such an inequality, even in this simple case, is still an active area of research. Most of these criterion depend on global geometric properties of the potential function, the simplest is given by the Bakry-Emery criterion. In the case of $\hat{\pi}$, this states that if R is smooth and ρ-strongly convex then $\hat{\pi}$ satisfies a logarithmic Sobolev inequality with constant ρ^{-1}. Here strong convexity refers to the uniform bound,

$$D^2 R \geq \rho I_d,$$

where D^2R denotes the Hessian of R (i.e. the eigenvalues of D^2R are bounded below by ρ). This, along with the previous analysis, suggests that the logarithmic Sobolev inequality can connect global geometric properties of the objective with rates of convergence and magnitude of excess risk. Of course, the potential of this connection is much more impressive in the non-convex case.

The approach taken in the original paper uses a combination of criterion that are referred to in the literature as *Lyapunov criterion*. These require the existence of a function $V \in C^2(\mathbb{R}^d)$ and constants $\lambda, \kappa, \delta > 0$ such that

$$\frac{\mathcal{L}V}{V} \leq -\lambda + \kappa \mathbb{1}_{B(0,\delta)},$$

where $\mathcal{L}$ is the infinitesimal generator (see appendix A) and $B(0,\delta)$ is the Euclidean ball of radius δ about the origin. Under such conditions a Poincaré inequality, which is much weaker, can be extended to a logarithmic Sobolev inequality. Additionally, an extremely crude estimate for a Poincaré inequality can be obtained. A discussion of how c_l is bounded in the paper is beyond the scope of these notes and we refer to appendices A, B and E of the original paper for the proof. A detailed review of criterion for such inequalities can be found in [1].

So far our analysis has depended on technical assumptions (B.1), (B.2) and (B.3) holding as well as $D(\mu_0\|\hat{\pi}) < \infty$. In the next proposition we show that such conditions can be replaced with neater sufficient conditions that are much easier to verify. These are the assumptions presented as part of the main result in the original paper.

Proposition 5.2. *Suppose the following holds:*

(A.1) There exists $M, A, B > 0$ such that for all $z \in \mathcal{Z}$, $f(\cdot, z)$ is M-smooth, $\|\nabla f(0, z)\| \leq B$ and $|f(0, z)| \leq A$.

(A.2) There exists $m > 0, b \geq 0$ such that for each $z \in \mathcal{Z}$, the function $f(\cdot, z)$ is (m, b)-dissipative: for all $w \in \mathbb{R}^d$, $\langle w, \nabla f(w, z)\rangle \geq m\|w\|^2 - b$.

(A.3) The probability law μ_0 of the initial hypothesis W_0 has a bounded and strictly positive density ρ_0 with respect to the Lebesgue measure on $\mathbb{R}^d$ and

$$\kappa_0 := \log \int_{\mathbb{R}^d} e^{\|w\|^2} \mu_0(dw) < \infty.$$

Then for any $\mathbf{z} \in \mathcal{Z}^n$ assumptions (B.1), (B.2) and (B.3) hold with $B_c = \kappa_0 + (b + d/\beta)/m$,

$$c_l = \mathcal{O}\big(\beta + d\big)^2 \exp(\mathcal{O}\big(\beta + d\big)\big), \tag{15}$$

and additionally $D(\mu_0\|\hat{\pi}) < \gamma$ with $\gamma = \mathcal{O}(\beta)$.

The proof of the relative entropy bound can be found in Lemma 3.4 of the original paper and the proof that (B.2) holds is given in Lemma 3.2.

Given a loss function f, (A.1) and (A.3) are certainly easy to verify, but assumption (A.2) may still seem somewhat unclear and overly-technical. A clearer sufficient condition is that for each $z \in \mathcal{Z}$ there exists $m_z \geq m$ such that

$$f(w, z) = m_z\|w\|^2 + f_0^z(w),$$

where f_0^z is a $2\sqrt{bm}$-Lipschitz continuous function. In other words, $f(\cdot, z)$ deviates from the quadratic function $m_z\|\cdot\|^2$ by a Lipschitz function.

5.3 Gaussian Smoothing for Fast Convergence

A popular approach for overcoming the difficulties of non-convex optimization is by smoothing the objective, often with a Gaussian convolution. This is referred to as global continuation.

Recent works [2, 10] have explored how taking Gaussian convolutions can sharpen bounds on logarithmic Sobolev constants in the compact case. In the next result, based on Theorem 1.2 of [2], we consider modifying the Langevin dynamics instead using

$$dW(t) = -\nabla \tilde{R}(w)dt + \sqrt{2\beta^{-1}}dB(t). \tag{16}$$

where $\tilde{R}$ is a particular smoothed form of R. The method of smoothing we consider has also appeared under the name Entropy-SGD [4].

Proposition 5.3. *Let $\sigma \geq 0$ and $\delta \geq \sigma$ and consider the smoothed empirical risk*

$$\tilde{R}(w) = -\frac{1}{\beta} \log \int_{\|v\| \leq \delta} \exp(-\beta R(w))\gamma_{\sigma^2}(v - w)dv.$$

Then the process defined by (16) has stationary distribution $\tilde{\pi} * \gamma_{\sigma^2}$ where

$$\tilde{\pi}(dw) = \tilde{\Lambda}^{-1} \exp(-\beta R(w))\mathbb{1}_{B(0,\delta)}(w)dw, \quad \tilde{\Lambda} = \int_{B(0,\delta)} \exp(-\beta R(w))dw.$$

which satisfies a logarithmic Sobolev inequality with constant

$$c_l = \left(K_1 d + K_2 \frac{\delta^2}{\sigma^2}\right)\delta^2 \exp\left(4\frac{\delta^2}{\sigma^2}\right),$$

where $K_1, K_2 > 0$ are universal constants.

Note that the exponential dependence on dimension has been traded for linear dependence which is highly desirable when applied to large models such as deep neural networks. However, this comes at the cost of having the objective constrained to a ball of finite radius and in practice, whatever computational expense comes with having to estimate the convolution. Though this result doesn't immediately point to a new algorithm it does suggest that the logarithmic Sobolev inequality could be a powerful tool for understanding global continuation and related methods such as artificial feature noise.

References

[1] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348 of *Grundlehren der mathematischen Wissenschaften*. Springer International Publishing, Cham, 2014. ISBN 978-3-319-00226-2. doi: 10.1007/978-3-319-00227-9. URL `http://link.springer.com/10.1007/978-3-319-00227-9`.

[2] Jean-Baptiste Bardet, Nathaël Gozlan, Florent Malrieu, and Pierre-André Zitt. Functional inequalities for Gaussian convolutions of compactly supported measures: explicit bounds and dimension dependence. *Bernoulli*, 24(1):333–353, 7 2015. URL `http://arxiv.org/abs/1507.02389`.

[3] V. S. Borkar and S. K. Mitter. A Strong Approximation Theorem for Stochastic Recursive Algorithms. *Journal of Optimization Theory and Applications*, 100(3):499–513, 1999. ISSN 00223239. doi: 10.1023/A:1022630321574. URL `https://link.springer.com/article/10.1023/A:1022630321574`.

[4] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, 11 2017. URL `https://github.com/ucla-vision/entropy-sgd`.

[5] Grigorios A. Pavliotis. *Stochastic Processes and Applications*, volume 60 of *Texts in Applied Mathematics*. Springer New York, New York, NY, 2014. ISBN 978-1-4939-1322-0. doi: 10.1007/978-1-4939-1323-7. URL `http://link.springer.com/10.1007/978-1-4939-1323-7`.

[6] Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-Convex Learning via Stochastic Gradient Langevin Dynamics: A Nonasymptotic Analysis. In *Proceedings of Machine Learning Research*, volume 65, pages 1–30. PMLR, 6 2017. URL `http://proceedings.mlr.press/v65/raginsky17a.html`.

[7] Ramon van Handel. Probability in High Dimension. Technical report, Princeton University, 6 2014. URL `https://apps.dtic.mil/sti/citations/ADA623999`.

[8] Cédric Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, Rhode Island, 3 2003. ISBN 9780821833124. doi: 10.1090/gsm/058. URL `http://www.ams.org/gsm/058`.

[9] M Welling and YW Teh. Bayesian learning via stochastic gradient langevin dynamics. *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, 2011.

[10] David Zimmermann. Logarithmic Sobolev inequalities for mollified compactly supported measures. *Journal of Functional Analysis*, 265(6):1064–1083, 9 2013. ISSN 00221236. doi: 10.1016/j.jfa.2013.05.029.

Appendix A Functional Inequalities for Diffusion Semigroups

Initially, these notes contained an overview of diffusion semigroups and an introduction to logarithmic Sobolev inequalities framed in full generality for this case. This was later swapped with restriction to the Langevin dynamics. For the interested reader, this is now presented in this appendix. Note that this is still an informal introduction, a more complete introduction to this is given in [1].

A.1 A Brief Overview of Diffusion Semigroups

Suppose $X(t)$ is a continuous-time Markov process on $\mathbb{R}^d$. Define the associated *Markov semigroup* $(P_t)_{t\geq 0}$ on suitable measurable functions φ by

$$P_t\varphi(x) = \mathbb{E}[\varphi(X(t))|X(0) = x].$$

It is referred to as a semigroup because it satisfies the semigroup property: $P_t \circ P_s = P_{t+s}$ for any $t, s \geq 0$ and $P_0 = \mathrm{Id}$, both of which follow directly from the Markov property of $X(t)$. As a result, the semigroup can be characterised entirely by its derivative at $t = 0$. We represent this derivative with the operator $\mathcal{L}$, which maps a function φ to the derivative of $P_t\varphi$ at $t = 0$. This operator is called the (infinitesimal) *generator* and is defined by

$$\mathcal{L}\varphi(x) = \lim_{t\to 0} \frac{P_t\varphi(x) - \varphi(x)}{t}.$$

Indeed, the *Kolmogorov equation* shows that $\mathcal{L}$ characterises the semigroup through the differential equation

$$\frac{d}{dt}P_t\varphi = P_t\mathcal{L}\varphi = \mathcal{L}P_t\varphi. \tag{17}$$

For each $t \geq 0$ we notate the distribution of $X(t)$ with ν_t. When studying the probability distribution associated with the semigroup it is helpful to use the L^2-adjoint of the generator, denoted $\mathcal{L}^*$ (i.e. $\langle \mathcal{L}f, g\rangle_{L^2} = \langle f, \mathcal{L}^*g\rangle_{L^2}$ where $\langle\cdot,\cdot\rangle_{L^2}$ denotes the L^2-inner product). Suppose ν_t has a density ρ_t for any $t \geq 0$ and $X(0)$ is deterministic. Then the right-hand side of (17) reads

$$P_t\mathcal{L}\varphi(X_0) = \int \mathcal{L}\varphi(x)\rho_t(x)dx = \langle \mathcal{L}\varphi, \rho_t\rangle_{L^2(\mathbb{R}^d)}.$$

On the other hand, we can obtain an alternative form for the derivative of P_t:

$$\partial_t P_t\varphi(X_0) = \frac{d}{dt}\int \varphi(x)\rho_t(x)dx = \langle \varphi, \partial_t\rho_t\rangle_{L^2(\mathbb{R}^d)}.$$

From equation (17) it is given that the previous two equations are equivalent. Thus, if φ is arbitrary in some suitably dense function space, it can be shown that

$$\partial_t\rho_t = \mathcal{L}^*\rho_t. \tag{18}$$

This result is referred to as the *Fokker-Planck equation* or *Kolmogorov forward equation*. In the case that $X(0)$ is non-deterministic an identical result is readily obtained.

Suppose ρ is a probability density with respect to the Lebesgue measure, then if it is the case that

$$\mathcal{L}^*\rho = 0$$

then ρ defines a *stationary* measure. Certainly, if $X_0 \sim \rho(d\lambda)$ then equation (18) yields $\rho_t = \rho$ for all $t \geq 0$. In some sense, ρ represents a state of equilibrium for $X(t)$.

In the case that $X(t) = W(t)$, the generator and its adjoint are given by

$$\begin{aligned} \mathcal{L}\varphi(x) &= -\nabla R(x)\cdot\nabla\varphi(x) + \beta^{-1}\nabla\cdot\nabla\varphi(x), \\ \mathcal{L}^*\varphi(x) &= \nabla\cdot\big(\nabla R(x)\varphi(x) + \beta^{-1}\nabla\varphi(x)\big). \end{aligned} \tag{19}$$

A simple computation reveals that $\mathcal{L}^*\hat{\pi} = 0$ and so $\hat{\pi}$ must be stationary. After finding a stationary measure it is natural to ask whether $X(t)$ converges towards it as $t \to \infty$, in some probabilistic sense. Behaviour of this sort is broadly referred to as *ergodicity*. In the next section we will focus on a form of ergodicity that is applicable to the non-asymptotic analysis.

In the interest of brevity we have completely avoided discussions about domains of definition for the operators defined above. Under the condition that $X(t)$ has a stationary measure μ we can show that the domains of $(P_t)_{t\geq 0}$ and $\mathcal{L}$, denoted $\mathcal{D}$ and $\mathcal{D}(\mathcal{L})$ respectively, are both dense in $L^2(\mu)$. In the case that $X(t)$ is a diffusion process (a solution to an SDE with *nice* coefficients) results that hold for bounded measurable functions extend naturally to the entire domain. A complete treatment of these domains is given in the first two chapters of [1].

A.2 Functional Inequalities and Exponential Ergodicity

We begin by defining one additional object for Markov processes with stationary measures. Suppose $X(t)$ has a stationary measure μ and define the associated *Dirichlet form*,

$$\mathcal{E}(\varphi) = -\langle \varphi, \mathcal{L}\varphi \rangle_{L^2(\mu)}.$$

The purpose of this object may seem unclear to begin with, but we will show that given the right function it can produce important statistical quantities. In the case that $X(t) = W(t)$ and $\mu = \hat{\pi}$ it can be shown using integration by parts that

$$\mathcal{E}(\varphi) = \beta^{-1}\hat{\pi}\big(\|\nabla\varphi\|^2\big). \tag{20}$$

Now we can state the most simple functional inequality. We say that μ satisfies a *Poincaré inequality* if for some $\lambda > 0$

$$\mathbb{E}_\mu|\varphi|^2 \leq \lambda^{-1}\mathcal{E}(\varphi),$$

for all suitable measurable φ. The reason this inequality is interesting becomes clearer once we set $\varphi = P_t\phi$ for some $\phi \in \mathcal{D}(\mathcal{L})$ and any $t \geq 0$. The Kolmogorov equation implies

$$\mathcal{E}(P_t\phi) = -\langle P_t\phi, \partial_t P_t\phi\rangle_{L^2(\mu)} = -\partial_t\mathbb{E}_\mu|P_t\phi|^2,$$

and hence with the Poincaré inequality we obtain $\partial_t\mathbb{E}_\mu|P_t\phi|^2 \leq -\lambda\mathbb{E}_\mu|P_t\phi|^2$ for all $t \geq 0$. By Gronwall's lemma, we obtain exponential decay in the second moment,

$$\mathbb{E}_\mu|P_t\varphi|^2 \leq e^{-\lambda t}\mathbb{E}_\mu|\varphi|^2.$$

In the case where $X(t) = W(t)$, we can see from equation (20) that the Poincaré inequality reduces to a property of the empirical risk R and the parameter β.

We say that μ satisfies a logarithmic Sobolev inequality with constant $c > 0$ if for any suitable real-valued function φ we have

$$Ent_\mu(\varphi^2) \leq 2c\mathcal{E}(\varphi).$$

Similar to the Poincaré inequality, we can apply the inequality with $\varphi = \sqrt{P_t\phi}$ where $\phi > 0$ to obtain exponential convergence in entropy

$$Ent_\mu(P_t\phi) \leq e^{-2t/c}Ent_\mu(\phi).$$

Let ν be any measure on $\mathbb{R}^d$ such that we have absolute continuity $\nu \ll \mu$ and set $\varphi = \mathrm{d}\nu/\mathrm{d}\mu$ to be the Radon-Nikodym derivative. Then the entropy of φ is given by the *relative entropy*

$$Ent_\mu(\varphi) = D(\nu\|\mu) = \int \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\log\frac{\mathrm{d}\nu}{\mathrm{d}\mu}d\mu.$$

Following from this, we state the main convergence result.

Theorem A.1. *Suppose $\mathcal{L}$ is self-adjoint on $L^2(\mu)$ and μ satisfies a logarithmic Sobolev inequality with constant $c > 0$. Then if $D(\nu_0\|\mu) < \infty$ we have exponential convergence in relative entropy:*

$$D(\nu_t\|\mu) \leq e^{-2t/c}D(\nu_0\|\mu).$$

Certainly in the case of $X(t) = W(t)$ a simple computation reveals that $\mathcal{L}$ is indeed self-adjoint. This condition is necessary to guarantee that $P_t(\mathrm{d}\nu_0/\mathrm{d}\mu) = \mathrm{d}\nu_t/\mathrm{d}\mu$.