Optimization, Isoperimetric Inequalities, and Sampling via Lyapunov Potentials

August Y. Chen AYC74@CORNELL.EDU
Cornell University, Department of Computer Science

Karthik Sridharan KS999@CORNELL.EDU
Cornell University, Department of Computer Science

Editors: Nika Haghtalab and Ankur Moitra

Abstract

In this paper, we prove that optimizability of any function F using Gradient Flow from all initializations implies a Poincaré Inequality for Gibbs measures $\mu_\beta \propto e^{-\beta F}$ at low temperature. In particular, under mild regularity assumptions on the convergence rate of Gradient Flow, we establish that μ_β satisfies a Poincaré Inequality with constant $O(\mathsf{C}_{\text{PI, LOCAL}})$ for $\beta \geq \Omega(d)$, where $\mathsf{C}_{\text{PI, LOCAL}}$ is the Poincaré constant of μ_β restricted to a neighborhood of the global minimizers of F. Under an additional mild condition on F, we show that μ_β satisfies a Log-Sobolev Inequality with constant $O(S\beta\mathsf{C}_{\text{PI, LOCAL}})$ where S denotes the second moment of μ_β. Here asymptotic notation hides F-dependent parameters. At a high level, this establishes that optimizability via Gradient Flow from every initialization implies a Poincaré and Log-Sobolev Inequality for the low-temperature Gibbs measure, which in turn imply sampling from all initializations.

Analogously, we establish that under the same assumptions, if F can be initialized from everywhere except some set $\mathcal{S}$, then μ_β satisfies a Weak Poincaré Inequality with parameters $(O(\mathsf{C}_{\text{PI, LOCAL}}), O(\mu_\beta(\mathcal{S})))$ for $\beta \geq \Omega(d)$. At a high level, this shows while optimizability from 'most' initializations implies a Weak Poincaré Inequality, which in turn implies sampling from suitable warm starts. Our regularity assumptions are mild and as a consequence, we show we can efficiently sample from several new natural and interesting classes of non-log-concave densities, an important setting with relatively few examples. As another corollary, we obtain efficient discrete-time sampling results for log-concave measures satisfying milder regularity conditions than smoothness, similar to Lehec (2023).

Keywords: Isoperimetric Inequalities, Non-log-concave sampling, Non-convex optimization, Langevin Dynamics, Markov Chain mixing time

1. Introduction

Sampling from a high-dimensional distribution is a fundamental algorithmic problem in Machine Learning (ML) and statistics, with several applications such as Bayesian inference (Gilks et al., 1995; Gamerman and Lopes, 2006; Stuart, 2010; Kroese et al., 2013; Chewi, 2024). Moreover, with the recent rise of generative AI methods such as diffusion models, this perspective on ML has become increasingly popular in practice; see e.g. Song and Ermon (2019); Ho et al. (2020); Song et al. (2021b,a). Recently, significant theoretical progress has been made in sampling from 'nice enough' – but still fairly general – distributions in $\mathbb{R}^d$ via the gradient-based Langevin Monte Carlo (LMC) method, which can be viewed as a natural variant of Gradient Flow (GF) and Gradient Descent (GD). It has recently been shown LMC can sample from the Gibbs measure $\mu_\beta = e^{-\beta F}/Z$, where Z denotes the partition function, F denotes the log-density or the *energy function*, and $\beta > 0$ is the inverse temperature, given access to a gradient oracle ∇F[1], if μ_β satisfies certain nice properties.[2]

1. Similar but weaker guarantees hold given access to a stochastic gradient oracle, which is not the focus of our work.
2. As with the rest of the literature on this topic, for the rest of the paper we assume the existence of μ_β for all $\beta \geq \Omega(1)$. Moreover, for the rest of the paper, we work in $\mathbb{R}^d$.

In continuous time, LMC is the *Langevin Diffusion*, the following Stochastic Differential Equation (SDE):

$$\mathrm{d}\mathbf{w}(t) = -\beta \nabla F(\mathbf{w}(t))\mathrm{d}t + \sqrt{2}\mathrm{d}\mathbf{B}(t). \tag{1}$$

Here $\mathbf{B}(t)$ denotes a standard Brownian motion in $\mathbb{R}^d$. This is a natural method to sample from μ_β: the continuous-time Langevin Diffusion with inverse temperature β, the SDE (1), converges to μ_β (Chiang et al., 1987). In discrete time, there are several discretizations of (1). One natural discretization is *Gradient Langevin Dynamics*, defined as follows:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta\beta\nabla F(\mathbf{w}_t) + \sqrt{2\eta}\boldsymbol{\varepsilon}_t. \tag{2}$$

Here $\eta > 0$ is the step size, $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, \boldsymbol{I}_d)$ is a d-dimensional standard Gaussian, and $\beta > 0$ is the *inverse temperature parameter* (when larger, noise is weighted less). Another is the *Proximal Sampler* which we elaborate on in Subsection B.2 (Lee et al., 2021; Chen et al., 2022; Liang and Chen, 2022a,b; Fan et al., 2023; Altschuler and Chewi, 2024). Yet another discrete-time sampler is the *Weakly Dissipative Tamed Unadjusted Langevin Algorithm* and the *Regularized Tamed Unadjusted Langevin Algorithm*, which we elaborate on in Subsection B.3 (Lytras and Mertikopoulos, 2024). Broadly, these algorithms are known as *Langevin Dynamics* and aim to discrete (1). Note as $\beta \to \infty$, reparametrizing (1) in terms of $t_{\text{new}} = \beta t$, (1) becomes GF with time t_{new}, and reparametrizing (2) in terms of $\eta_{\text{new}} = \eta\beta$, (2) becomes GD with step size η_{new}.

It is now established that continuous and discrete time LMC can sample from μ_β *beyond log-concavity* (when F is convex), to when μ_β satisfies an *isoperimetric inequality*, which correspond to geometric properties of F allowing the Markov process (1) to mix efficiently (Villani, 2009, 2021; Bakry et al., 2014).

- The most general such inequality under which discrete-time LMC has been proved to be successful from *arbitrary* initialization is when μ_β satisfies a *Poincaré Inequality* (PI) (Chewi et al., 2024).
- A stronger, related inequality under which discrete-time LMC efficiently succeeds is when μ_β satisfies a *Log-Sobolev Inequality* (LSI) (Vempala and Wibisono, 2019). This is referred to as the 'sampling analogue of gradient domination', as it implies gradient domination in Wasserstein space (Jordan et al., 1998).
- Under a *Weak Poincaré Inequality* (WPI), which generalizes a PI, continuous time LMC can efficiently sample from μ_β from a suitable *warm start* (Röckner and Wang, 2001; Wang, 2006; Bakry et al., 2014; Mousavi-Hosseini et al., 2023; Huang et al., 2025).

We defer more discussion on isoperimetric inequalities to Subsection 2.1. Such sampling results have in turn been used to show appropriately-scaled LMC can optimize non-convex F to tolerance $\tilde{O}\left(\frac{d}{\beta}\right)$ (Raginsky et al., 2017; Xu et al., 2018; Zou et al., 2021).[3]

However, it is not clear what this means more concretely. Classically, when F is convex, μ_β satisfies a PI (Bobkov, 1999); when F is strongly convex, μ_β satisfies a LSI (Bakry and Émery, 2006). But beyond convexity, do we have classes of energy functions/log-densities F for which μ_β satisfies isoperimetry? For example, when F satisfies gradient domination in the traditional sense of optimization – which allows for GF and GD to optimize F – does μ_β satisfy a PI or LSI (and consequently we can sample from it)?

Before highlighting our results, we mention that related works and a comparison to our results, including the concurrent works Chewi and Stromme (2024); Gong et al. (2024), can be found next in Subsection 1.2.

Convention. For the rest of paper, by shifting we assume WLOG that F attains a minimum value of 0 on $\mathbb{R}^d$. We let $\mathbf{w}^\star$ denote any arbitrary global minimizer of F, thus $F(\mathbf{w}^\star) = 0$.

3. In runtime worst-case exponential in β.

1.1. Overview of Results

PI/LSI Results: The similarity between Langevin Dynamics and GF/GD motivates the overarching:

Conjecture 1 *If F is optimizable via Gradient Descent from arbitrary initialization, then $\mu_\beta := e^{-\beta F}/Z$ satisfies a PI for appropriate β. Thus we can efficiently sample from μ_β for such β with oracle access to ∇F.*

This is natural: if gradient-based methods succeed for optimization without getting stuck, LMC ought to not get stuck as well. Moreover, ∇F is the exact same oracle as we have for GF/GD.

We proceed to define optimizability of F via GF following Definition 1 and Theorem 2 of De Sa et al. (2022). This following condition is derived in De Sa et al. (2022) from the existence of a appropriate rate function for the convergence of GF. The notion of appropriate rate function from De Sa et al. (2022) is very generic – for example, is is satisfied whenever GF enjoys an exponential rate – and as such the following definition covers numerous examples in non-convex optimization. See Section 4 for a subset of these examples.

Definition 2 (Optimizability of F via Gradient Flow) *For F with minimum value 0, we say F is optimizable by Gradient Flow if for all $\mathbf{w} \in \mathbb{R}^d$, there exists a Lyapunov Function $\Phi(\cdot)$ such that*

$$\langle \nabla \Phi(\mathbf{w}), \nabla F(\mathbf{w}) \rangle \geq g(F(\mathbf{w})), \tag{3}$$

where g is monotonically non-decreasing with $g(0) = 0$ and $g(x) > 0$, $g(x) \geq m'x - b'$ for some $m', b' > 0$ for all $x > 0$.[4] *Moreover, we say F is optimizable by GF from a set $\mathcal{Q} \subset \mathbb{R}^d$ if (3) holds for all $\mathbf{w} \in \mathcal{Q}$.*

Convention. We simply refer to F as optimizable when F is optimizable by GF in the sense of Definition 2.

Moreover, to obtain a PI and therefore discrete-time sampling results, it is natural to assume discrete-time optimization via GD in addition to GF succeeds. For GD to succeed in optimizing F (i.e. for Taylor terms in GD to be controlled), we require that Φ and F satisfy the following assumption:

Assumption 1.1 (Self-Bounding Regularity) *For some monotonically non-decreasing $\rho_\Phi, \rho_F : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, we have $\|\nabla \Phi(\mathbf{w})\|, \|\nabla^2 \Phi(\mathbf{w})\|_{\text{op}} \leq \rho_\Phi(\Phi(\mathbf{w}))$ and $\|\nabla F(\mathbf{w})\|, \|\nabla^2 F(\mathbf{w})\|_{\text{op}} \leq \rho_F(F(\mathbf{w}))$.*[5]

As shown in Theorem 3 of De Sa et al. (2022), assumptions analogous to Assumption 1.1 are actually *necessary* for GD to succeed for discrete-time optimization, and hence come with little loss of generality. Note smoothness of Φ and F (e.g. $\Phi = F$ for PŁ functions) is a special case of Assumption 1.1, but Assumption 1.1 is much more general. Such a framework with dimension-independent ρ_Φ, ρ_F subsumes numerous examples in non-convex (and convex) optimization; see Section 4 and De Sa et al. (2022).

We confirm Conjecture 1 in the following sense, stated formally in Theorem 12. Under Assumption 1.1, Assumption 3.1 (which subsumes the literature and is necessary, see Remark 11), and Assumption 3.2:

$$\text{Optimizability of } F \text{ for all } \mathbf{w}, \text{ i.e. (3)} \implies \text{PI for } \mu_\beta \text{ for } \beta = \Omega(d) \text{ with PI constant } O(\text{poly}(d, \beta)). \tag{4}$$

In Theorem 12, we furthermore establish:

$$\text{Above conditions + mild regularity on } F \implies \text{LSI for } \mu_\beta \text{ for } \beta = \Omega(d) \text{ with LSI constant } O(\text{poly}(d, \beta)).$$

In comparing optimization to sampling for F optimizable by GF/GD, $\beta = \Omega(d)$ is the correct scaling (and has several applications); see Subsection 2.2. When $\beta = \Omega(d)$ is written above, the asymptotic notation hides F-dependent constants; see e.g. Remark 13 and Subsection D.1 for full expressions. As a direct consequence

4. We assume $g(x)$ has at least linear tail growth, as g arises to handle when the rate function $R(\mathbf{w}, t)$ for GF is not integrable, e.g. for convex rate t^{-1}.

5. In fact the bound on operator norm implies the bound on the gradient; see Lemma 11, De Sa et al. (2022).

of the literature, having established a PI and/or LSI, we obtain that discrete-time LMC can sample from μ_β for such β in time polynomial in $d, \beta, \frac{1}{\varepsilon}$ under very mild regularity assumptions; see Corollary 19, Corollary 20.

We view this as a core strength of our work: our result complements the literature and 'plugs and plays' with sampling algorithms and their analysis that study sampling under isoperimetry. We further emphasize that *the focus of our work is not to develop or analyze sampling algorithms, but rather to prove that geometric properties imply functional inequalities (PI/WPI)*, which are the *crux* of LMC. To obtain Corollary 19, Corollary 20 we simply take results in the literature that, to the best of our knowledge, have the state-of-the-art results for LMC.

For these corollaries we make no warm start assumption, and instead explicitly describe the initialization, which does not depend on $\mathbf{w}^\star$. Our sampling algorithms succeed solely because F is optimizable everywhere; intuitively, LMC 'moves' us towards μ_β due to the optimizability condition $\langle \nabla \Phi(\mathbf{w}), \nabla F(\mathbf{w}) \rangle \geq g(F(\mathbf{w}))$. If optimizability only holds within $\mathbb{B}(\mathbf{w}^\star, R)$ for some R, we show in Remark 18 (with details in Proposition 21) that by appropriately regularizing on F outside $\mathbb{B}(\mathbf{w}^\star, R)$ to yield $\hat{F}$, we can sample from $\hat{\mu}_\beta \propto \exp(-\beta \hat{F}) \approx \mu_\beta$ (the approximation holds for R large). We view this as an interesting algorithmic implication of our work.

Weak Poincaré Inequalities: In many non-convex landscapes, such as Phase Retrieval and Matrix Square Root, there is a set $\mathcal{S}$ with small Lebesgue measure of bad initializations where GF/GD does not succeed, but everywhere else GF/GD works (Jain et al., 2017; Lee et al., 2019; De Sa et al., 2022). It can be moreover verified that outside $\mathcal{S}$, optimizability as per Definition 2 holds (De Sa et al., 2022). Little is known about sampling in such settings. As such a deeper understanding of these settings is very important and interesting.

A Weak Poincaré Inequality (WPI) captures this picture, corresponding to efficient sampling under a *warm start* which has low density in $\mathcal{S}$. It is crucial to note such a situation is *not covered by a PI*, as a PI implies *worst-case* mixing. Thus it is natural to expect:

Conjecture 3 *If F is optimizable via Gradient Descent from everywhere except a set $\mathcal{S}$ with small Lebesgue measure, then μ_β satisfies a $(\mathsf{C}_{\text{WPI}}, \delta)$-WPI with δ small for appropriate β. (See Subsection 2.1 for the formal definition of a WPI; here δ in the WPI controls the 'error' we can sample to efficiently.) Thus we can efficiently sample from μ_β for such β with oracle access to ∇F with a warm start.*

For clarity on what we mean by F being optimizable via Gradient Descent from everywhere except a set $\mathcal{S}$ with small Lebesgue measure, we mean that for all $\mathbf{w} \in \mathbb{R}^d \setminus \mathcal{S}$, (3) holds. That is, we have for some Φ and g satisfying the conditions of Definition 2,

$$\langle \nabla \Phi(\mathbf{w}), \nabla F(\mathbf{w}) \rangle \geq g(F(\mathbf{w})) \text{ for all } \mathbf{w} \in \mathbb{R}^d \setminus \mathcal{S}.$$

We denote this by 'optimizability of F from $\mathcal{S}^c$'. As a concrete example, this holds if F is PŁ outside of some $\mathcal{S} \subset \mathbb{R}^d$.

Indeed, we confirm Conjecture 3 in the following sense, stated formally in Theorem 15. We show under Assumption 1.1, Assumption 3.1, Assumption 3.2 that

$$\text{Optimizability of } F \text{ from } \mathcal{S}^c \implies (\mathsf{C}_{\text{WPI}}, O(\mu_\beta(\mathcal{S})))\text{-WPI for } \mu_\beta,\ \beta = \Omega(d),\ \mathsf{C}_{\text{WPI}} \approx \mathsf{C}_{\text{PI}} \text{ from (4)}. \quad (5)$$

Thus if $\mu_\beta(\mathcal{S})$ is small (e.g. if $\mathcal{S}$ has small Lebesgue measure and $\inf_{\mathbf{w} \in \mathcal{S}} F(\mathbf{w})$ is not too small), the above shows we can sample to low error via LMC from a warm start. Again here, the $O(\cdot)$, $\Omega(\cdot)$ hide F-dependent parameters. With a WPI, sampling from a warm start follows via e.g. Röckner and Wang (2001); Mousavi-Hosseini et al. (2023); Huang et al. (2025). Note $\mathcal{S}$ is arbitrary; it can comprise of saddle points or even spurious local minima.

Applications and Significance: Our results Theorem 12, Theorem 15 yield a natural, novel host of non-log-concave measures $\mu_\beta \propto \exp\{-\beta F(\mathbf{w})\}$ where LMC samples in time poly$(d, \beta, 1/\varepsilon)$. The crux to establishing these polynomial guarantees is showing poly(d, β) bounds on isoperimetric constants. This is known for log-concave measures (convex F); beyond log-concavity, known methods (e.g. perturbation criteria) give $\exp(d)$ constants, even for e.g. PŁ F. By contrast, Definition 2, and hence our results, subsumes the following general non-convex function classes for which GF/GD succeed for global optimization: Polyak-Łojasiewicz (PŁ) (Polyak, 1963; Lojasiewicz, 1963), Kurdyka-Łojasiewicz (KŁ) (Kurdyka, 1998), and Linearizable (Kale et al., 2021) functions (also known as Quasar-Convexity (Hinder et al., 2020)).

Definition 4 (Polyak-Łojasiewicz (PŁ)) *A differentiable function F is Polyak-Łojasiewicz (PŁ) with parameter $\lambda > 0$ if $\|\nabla F(\mathbf{w})\|^2 \geq \lambda F(\mathbf{w})$ for all $\mathbf{w} \in \mathbb{R}^d$. (Take $\Phi = F$, $g(x) = \lambda x$ in Definition 2. Recall we shifted so F has minimum value 0 before this section.)*

Definition 5 (Kurdyka-Łojasiewicz (KŁ)) *A differentiable function F is Kurdyka-Łojasiewicz (KŁ) with parameter $\lambda > 0$, $\theta \in [0, 1)$ if $\|\nabla F(\mathbf{w})\|^2 \geq \lambda F(\mathbf{w})^{1+\theta}$ for all $\mathbf{w} \in \mathbb{R}^d$. (Take $\Phi = F$, $g(x) = \lambda x^{1+\theta}$.)*

Definition 6 (Linearizable) *A differentiable function F is λ-linearizable if for some global minimizer $\mathbf{w}^\star \in \mathbb{R}^d$ of F, $\langle \nabla F(\mathbf{w}), \mathbf{w} - \mathbf{w}^\star \rangle \geq \lambda F(\mathbf{w})$ for all $\mathbf{w} \in \mathbb{R}^d$.[6] (Take $\Phi = \|\mathbf{w} - \mathbf{w}^\star\|^2$, $g(x) = \lambda x$.)*

Consequently Theorem 12 yields a PI and thus poly$(d, \beta, 1/\varepsilon)$ sampling guarantees for $\mu_\beta \propto \exp(-\beta F)$ when F is in the above classes, under the conditions of Theorem 12. Analogously, under the conditions of Theorem 15, we obtain a WPI and sampling from a warm start for $\mu_\beta \propto \exp(-\beta F)$ when F is in the above classes. Such a result has further applications and interpretation in Bayesian inference, as we detail in Subsection 2.2. This alternate interpretation is: if Maximum a posteriori is optimizable by GF/GD, LMC efficiently samples from the posterior. Optimizability of ERM/regularized ERM by GF/GD has been studied extensively; our work lets us systematically use such results to fuel Bayesian inference.

For another application, note general convex functions are 1-Linearizable and automatically satisfy Assumption 3.1. Corollary 23 thus gives Corollary 24, a poly$\left(\beta, d, \frac{1}{\varepsilon}\right)$ sampling guarantee for log-concave measures at low temperatures under relaxed regularity assumptions (beyond smoothness). Log-concave sampling beyond smoothness was studied in Lehec (2023); our regularity assumptions are in some sense more general.

Technical Approach: We also highlight our technical approach. We utilize this exact Lyapunov function Φ from optimization (from Definition 2) to execute the Lyapunov potential technique from probability (Bakry et al., 2008) to prove a PI/LSI. Generally the technique of Bakry et al. (2008) involves significantly different Lyapunov potentials than those from optimization, and often ad-hoc. Using the exact same potential from optimization gives crisp quantitative control over the isoperimetric constants of μ_β. This crisp quantitative control stands in contrast to typical usages of this technique. We also further develop this technique to prove a WPI. To the best of our knowledge, our work is the first to develop the Lyapunov function technique to establish a WPI. Our means of using the Lyapunov function technique to establish a WPI is simple and user-friendly, and we expect that it will have further applications. *As such, our work tightens the link between optimization, sampling, and probability in several ways.*

Connecting Optimization and Sampling: Our results yield *fundamental relationships at the algorithmic level, connecting optimizability via GF/GD to isoperimetry at low temperature (and hence the success of Langevin Dynamics).* There are several connections between sampling and optimization, from the Proximal Point Method of optimization inspiring the Proximal Sampler, to interior point methods for log-concave sampling (Kook and Vempala, 2024). Here, we address Conjecture 1, Conjecture 3 and deepen the connection between optimization, isoperimetry, and sampling from another angle.

6. We make a change of variables compared to its definition in (Kale et al., 2021).

1.2. Related Works

Several other works have studied the connection between efficient optimization, isoperimetry, and sampling:

- Ma et al. (2019) studied this connection across *different* temperature levels β, where the behavior of μ_β fundamentally changes. In contrast, we study a given, fixed landscape for large β, and study the connection between optimization and sampling in this landscape.

- Several recent works (Li and Erdogdu, 2023; Kinoshita and Suzuki, 2022; Lytras and Sabanis, 2023; Huang and Sellke, 2025+; Sellke, 2024) show that when the landscape of $-\log \mu_\beta = F$ is strict saddle in the sense of a constant order negative eigenvalue around spurious critical points, then combined with several other regularity assumptions, functional inequalities hold. Among these, Kinoshita and Suzuki (2022); Lytras and Sabanis (2023) studies the problem in Euclidean spaces. However, this does not capture our setting of general functions optimizable by GF/GD. Thus these settings are not comparable. Indeed, there are many functions where GF/GD succeed that are not strict-saddle, such as star-convex functions, smooth one-point-strongly convex functions, and even general convex functions. See Example 5 for further discussion of these examples. These results also contain many unnecessary regularity assumptions and/or suboptimal F-dependent parameters. We bypass these suboptimal dependencies via our novel use of the Lyapunov function method.

  Moreover, the results of Kinoshita and Suzuki (2022); Lytras and Sabanis (2023); Li and Erdogdu (2023) only hold for an unreasonably low temperature regime, $\beta \geq \Omega(d^6)$, where $\Omega(\cdot)$ again hides F-dependent parameters. This is often much larger than $\beta = \tilde{\Theta}\left(\frac{d}{\varepsilon}\right)$ used for optimization via LMC to tolerance ε. At such the algorithmic implications of their result simply is that optimization is possible in strict-saddle landscapes. By contrast, this is *not* the case for $\beta \geq \Omega(d)$ as we consider, the regime where sampling is of interesting in Bayesian inference; see Subsection 2.2.

- The concurrent works Chewi and Stromme (2024); Gong et al. (2024) study a special case of our problem, when F is PŁ and β is large (a setting subsumed by our Theorem 12), also proceeding through Lyapunov functions.

  Gong et al. (2024) studies this problem under a local PŁ condition around local minima. However, they place several regularity assumptions on all of $\mathbb{R}^d$, which in they show in their Proposition 3.1 in fact imply unimodality analogous to our setting. Their Proposition 3.1 implies the existence of a connected set of local minima (see their note on page 3) and no saddle points. They further require a strictly negative lower bound on the Laplacian ΔF when the gradient is small, which factors into their quantitative dependencies; furthermore, such a situation can handled by our exact same proof, see Section C. Thus their work reduces to a setting analogous to ours. Their bound on the PI constant also implicitly incurs exponential d dependence; it contains a term of the form $\exp(\overline{C})$ (their Theorem 2, Lemma 4), and $\overline{C} = \Omega(M_\Delta) = \Omega(d)$ for generic smooth F (Lemmas 2, 3).[7]

  Chewi and Stromme (2024) obtains a sharp characterization of the Poincaré and Log-Sobolev constants of μ_β when F is PŁ and has a unique minimizer $\mathbf{w}^\star$ in the *asymptotic* limit $\beta \to \infty$. In this asymptotic limit, sampling degenerates into optimization and consequently the algorithmic implications of their result is relatively limited.[8] We also remark that our Theorem 12 implies their upper bound on the Poincaré constant up to a universal constant factor of 2 (see e.g. Remark 13), and that a Poincaré Inequality is sufficient to give an efficient sampling algorithm (see for instance Chewi et al. (2024); Lytras and Mertikopoulos (2024)).

7. Note they adopt convention that smaller PI constant is worse.
8. We point out their result will only hold for $\beta \geq \Omega(d)$ where asymptotic notation hides F-dependent parameters, since they require an upper bound on the Laplacian of F, which scales with d even for e.g. quadratics.

By contrast, our general optimizability condition is far more comprehensive and allows us to capture many examples under a single umbrella. It captures not only PŁ but also KŁ, Linearizable, Star-Convex, One-Point-Convex and general convex functions (see Example 4, Example 5). As an extreme example, convex F need not be PŁ, but are readily subsumed by our setting (see Example 5).

Our method of using Lyapunov functions is also novel, in that we prove functional inequalities using the *same* Lyapunov function arising from optimization, further highlighting the connection between optimization and sampling. Our techniques also yield improved quantitative dependencies on F-dependent parameters; see Remark 32. As a consequence of our general optimizability condition, beyond a wide host of applications (Example 3, Example 4, Example 5), *we obtain fundamental relationships at the algorithmic level: that optimizability, at appropriate β, implies the success of Langevin Dynamics for sampling.*

Furthermore, none of these works connect optimizability outside of some unfavorable region $\mathcal{S}$ (as is often the case in non-convex landscapes, e.g. Phase Retrieval) to a WPI, as we do in Theorem 15. Gong et al. (2024) allows for local maxima outside a local region (which as remarked above can be readily handled by our proof), but do not permit saddle points or spurious local minima as we do in Theorem 15. We also present algorithm implications of our result via regularization if we only have 'local' optimizability in Proposition 21 but arbitrary stationary points/spurious local minima elsewhere, a perspective unexplored in these works.

2. Preliminaries and Technical Background

2.1. Isoperimetric Inequalities

Isoperimetric inequalities define geometric properties of F that enable LMC (or other Markov Chains, which we do not expand on here) to mix rapidly. These isoperimetric inequalities are governed by their *isoperimetric constant*; in this work we adopt the notion that a smaller isoperimetric constant implies a stronger inequality. From *arbitrary* initializations, the most general condition under which LMC is successful is when μ_β satisfies a *Poincaré Inequality* (PI) (Villani, 2021; Bakry et al., 2014), defined as follows:

Definition 7 (Poincaré Inequality (PI)) *A probability measure μ on $\mathbb{R}^d$ satisfies a Poincaré Inequality (PI) with constant $\mathsf{C}_{\text{PI}}(\mu)$ if for all infinitely differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$, we have*

$$\int_{\mathbb{R}^d} f^2 \mathrm{d}\mu - \left(\int_{\mathbb{R}^d} f \mathrm{d}\mu\right)^2 \leq \mathsf{C}_{\text{PI}}(\mu) \int_{\mathbb{R}^d} \|\nabla f\|^2 \mathrm{d}\mu.$$

If the above is not satisfied, following the convention, we set $\mathsf{C}_{\text{PI}}(\mu) = \infty$.

A PI corresponds to exponential contraction of variance for the Langevin Diffusion (1) (note the left hand side can be written as the variance $\mathbb{V}_\mu(f)$), and directly implies continuous-time sampling results in χ^2-divergence via Langevin Dynamics (1). In particular, letting π_T denote the probability measure obtained after running the Langevin Diffusion (1) (with $-\log \mu$ in place of $\beta \nabla F$) for time T and π_0 denote the initialization, we have

$$\chi^2(\pi_T \| \mu) \leq e^{-2T/\mathsf{C}_{\text{PI}}(\mu)} \chi^2(\pi_0 \| \mu).$$

For both of these results, see e.g. Chapter 4, Bakry et al. (2014). By Bobkov (1999), if μ is *log-concave*, or equivalently $-\log \mu$ is a convex function of $\mathbf{w}$, then μ_β satisfies a PI. We next define *Log-Sobolev Inequality (LSI)*, which is stronger than PI.

Definition 8 (Log-Sobolev Inequality (LSI)) *A probability measure μ on $\mathbb{R}^d$ satisfies a Log-Sobolev Inequality (LSI) with Log-Sobolev constant $\mathsf{C}_{\text{LSI}}(\mu)$ if for all infinitely differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$, we have*

$$\int_{\mathbb{R}^d} f \ln f \mathrm{d}\mu - \int_{\mathbb{R}^d} f \ln\left(\int_{\mathbb{R}^d} f \mathrm{d}\mu\right) \mathrm{d}\mu \leq 2\mathsf{C}_{\text{LSI}}(\mu) \int_{\mathbb{R}^d} \|\nabla f\|^2 \mathrm{d}\mu.$$

If the above is not satisfied, following the convention, we set $\mathsf{C}_{\text{LSI}}(\mu) = \infty$.

A LSI has been referred to as the 'sampling analogue of the PŁ Inequality', since it implies gradient domination in Wasserstein space (Chewi, 2024). A LSI corresponds to exponential contraction of entropy $\text{ent}_\mu(f)$ for the Langevin Diffusion (1), which again is the left hand side of the above, and directly implies exponential contraction for the KL-divergence via the Langevin Diffusion (1) (run with $-\log\mu$ in place of $\beta\nabla F$). Namely defining π_T, π_0 as earlier, a LSI implies

$$\mathsf{KL}(\pi_T\|\mu) \le e^{-2T/\mathsf{C}_{\text{LSI}}(\mu)}\mathsf{KL}(\pi_0\|\mu).$$

See e.g. Chapter 5, Bakry et al. (2014). A LSI is stronger than a PI with the same constant: a LSI with constant $\mathsf{C}_{\text{LSI}}(\mu)$ implies that a PI with the same constant holds, thus $\mathsf{C}_{\text{PI}}(\mu) \le \mathsf{C}_{\text{LSI}}(\mu)$, but the reverse implication does not hold (Chewi, 2024). Obtaining a sampling result in KL (via LSI) is also stronger than in χ^2 (via PI). Indeed, not all log-concave measures satisfy a LSI.

From a suitable *warm-start*, the Langevin Diffusion can efficiently sample from μ_β under a *Weak Poincaré Inequality* (WPI) (Röckner and Wang, 2001; Wang, 2006; Bakry et al., 2014; Mousavi-Hosseini et al., 2023; Huang et al., 2025), which captures *beyond worst-case mixing*. Consider e.g. a mixture of two well-separated identity covariance Gaussians: mixing from arbitrary initialization is exponentially slow in d, but starting from a normal perfectly centered between the modes, we could conceivably obtain rapid mixing. Indeed, several works in probability have studied sampling from complicated distributions satisfying a WPI by 'chaining together' warm starts (Alaoui et al., 2025+; Huang et al., 2025). To define a WPI, we adopt convention from Definition 4.7, Huang et al. (2025).[9]

Definition 9 (Weak Poincaré Inequality (WPI)) *A probability measure* μ *on* $\mathbb{R}^d$ *satisfies a* $(\mathsf{C}_{\text{WPI}}(\mu), \delta)$*-Weak Poincaré Inequality (WPI) if for all infinitely differentiable functions* $f : \mathbb{R}^d \to \mathbb{R}$*, letting* $osc(f) = \sup f - \inf f$*, we have*

$$\int_{\mathbb{R}^d} f^2 \mathrm{d}\mu - \left(\int_{\mathbb{R}^d} f \mathrm{d}\mu\right)^2 \le \mathsf{C}_{\text{WPI}}(\mu) \int_{\mathbb{R}^d} \|\nabla f\|^2 \mathrm{d}\mu + \delta \text{osc}(f)^2.$$

Note $\text{osc}(f) \le 2\sup(|f - \mathbb{E}[f]|)$, so applying Theorem 2.1 of Röckner and Wang (2001) as in (2) of Huang et al. (2025) and defining π_T, π_0 as earlier, we have the following mixing guarantee for the continuous-time Langevin Diffusion (1) (again, run with $-\log\mu$ in place of $\beta\nabla F$)):

$$\chi^2(\pi_T\|\mu) \le e^{-T/\mathsf{C}_{\text{WPI}}(\mu)}\chi^2(\pi_0\|\mu) + 4\delta\left\|\frac{\mathrm{d}\pi_0}{\mathrm{d}\mu} - 1\right\|_\infty^2. \tag{6}$$

Thus if π_0 is a suitable warm start in that $\left\|\frac{\mathrm{d}\pi_0}{\mathrm{d}\mu} - 1\right\|_\infty^2$ is small, then we obtain a mixing guarantee. Hence δ is the 'error' or 'slack' in the WPI, indicating how accurately we can sample efficiently with a warm start. Thus in Theorem 15, if $\mu_\beta(\mathcal{S})$ is small, we can sample efficiently in continuous-time to small accuracy.

It is also worth discussing the tail growth of F for which $\mu_\beta = e^{-\beta F}/Z$ satisfies an isoperimetric inequality (Chewi et al., 2024; Mousavi-Hosseini et al., 2023). A PI for μ_β goes hand-in-hand with F having at least *linear* tail growth (e.g. $F(\mathbf{w}) = \|\mathbf{w}\|$). For example, we can prove F has linear tail growth if F is convex and μ_β exists; see Lemma 2.2, Bakry et al. (2008). A LSI for μ_β goes hand-in-hand with F having at least *quadratic* tail growth (e.g. $F(\mathbf{w}) = \|\mathbf{w}\|^2$). As such, it is natural to assume that F has linear tail growth to prove a PI, and that F has quadratic tail growth to prove a LSI.

9. The definition above in fact implies Definition 4.7 of Huang et al. (2025).

2.2. The Role of Temperature and Applications of Low-Temperature Sampling

Notice in our earlier results that the inverse temperature $\beta = \Omega(d)$. Justification for this scaling to study the connection between optimization and sampling is severalfold:

- Optimization is fundamentally performed at low temperature. Even at the initialization of optimization algorithms, the value of F at initialization is often viewed as $O(1)$ in the literature (De Sa et al., 2022; Bubeck et al., 2015; Nesterov et al., 2018), which corresponds to the inverse temperature $\beta = \Omega(d)$; consider initializing at $\mathcal{N}(\vec{\mathbf{0}}, \frac{1}{\beta}\boldsymbol{I}_d)$. Furthermore the temperatures range we consider corresponds to *initialization* ($\beta = \Omega(d)$) rather than *output* of optimization to tolerance ε ($\beta = \Omega(d/\varepsilon)$).
- We use $\beta = \Omega(d)$ simply to follow the above aforementioned scaling from optimization. It is possible to obtain an analogous result to ours in the $\beta = O(1)$ setting by changing Assumption 3.1 so that $\operatorname{diam}(\mathcal{W}^\star), r(l_b) = \Theta(\sqrt{d})$ rather than $\Theta(1)$ and $l_b = \Omega(\sqrt{d})$. Such a scaling is made for instance in Huang and Sellke (2025+). Then one can simply follow the same proof as ours from Section C.

Sampling at low temperature is also of independent theoretical interest and has been studied in several works, discussed in Subsection 1.2. Typically one expects that as β increases, the isoperimetric constants of μ_β become larger, or isoperimetric inequalities break altogether. This behavior has been rigorously confirmed in measures from statistical physics (El Alaoui and Gaitonde, 2024). As we establish here, such behavior sharply contrasts to when F is optimizable, despite the lack of global convexity. Note under the generality here, finding the exact temperature threshold for β where a PI/LSI holds is likely extremely difficult.[10]

Applications of Low-Temperature Sampling: Consider posterior sampling in high-dimensional regression. Following Montanari and Wu (2023) (who use p rather than d), we observe covariates $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and response $\boldsymbol{y}_0 \in \mathbb{R}^n$, under the linear model $\boldsymbol{y}_0 = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\vec{\mathbf{0}}, \sigma^2\boldsymbol{I}_n)$. We consider the proportional asymptotics $n/d \to \delta \in (0,\infty)$, common in high-dimensional statistics (see e.g. Barbier et al. (2018); Wainwright (2019)). Here, $\hat{F}_{\text{erm}}(\boldsymbol{\theta}) = \frac{1}{n}\|\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{\theta}\|_2^2$ is empirical risk approximating population loss. Write the product measure $\pi_\Theta^{\otimes d}(\mathrm{d}\boldsymbol{\theta}) \propto \exp(-R(\boldsymbol{\theta}))$. By eq. 12, p. 6 of Montanari and Wu (2023), the desired posterior to sample from is $\mu_{\boldsymbol{X},\boldsymbol{y}_0}(\mathrm{d}\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y}_0 - \boldsymbol{X}\boldsymbol{\theta}\|_2^2\right)\pi_\Theta^{\otimes d}(\mathrm{d}\boldsymbol{\theta})$. Thus as $n/d \to \delta$,

$$\mu_{\boldsymbol{X},\boldsymbol{y}_0}(\mathrm{d}\boldsymbol{\theta}) \propto \exp\left\{-\frac{d\delta}{2\sigma^2}\left(\hat{F}_{\text{erm}}(\boldsymbol{\theta}) + \frac{2\sigma^2}{n}R(\boldsymbol{\theta})\right)\right\}.$$

This posterior is in the low temperature regime we consider, i.e. $\beta = d\cdot\frac{\delta}{2\sigma^2} = \Omega(d)$. The objective we consider is the regularized ERM objective. Under the same asymptotics, the posterior is at low temperature $\beta = \Omega(d)$ more generally: for a Gibbs prior $\propto \exp(-R(\boldsymbol{\theta}))$ for $\boldsymbol{\theta}$ and the model $(\boldsymbol{y}_0)_i = f(\boldsymbol{x}_i;\boldsymbol{\theta}) + \boldsymbol{\varepsilon}_i \forall 1 \le i \le n$. For example, GLMs, where $f(\boldsymbol{x}_i;\boldsymbol{\theta}) = \psi(\boldsymbol{x}_i^\top\boldsymbol{\theta})$, are non-convex but are optimizable by GF/GD under appropriate conditions on $\psi(\cdot)$ (Foster et al., 2018; Wang and Wibisono, 2023). Note our results apply when $\beta/d = \Theta(1)$ as $d \to \infty$; in the regime $\beta/d \to \infty$ as $d \to \infty$, as is the case for several related works mentioned in Subsection 1.2, posterior sampling is significantly easier (Montanari and Wu, 2023; Bontemps, 2011).

One can interpret our result from a Bayesian inference perspective: if Maximum a posteriori is optimizable by GF/GD, LMC efficiently samples from the posterior. Optimizability of ERM and regularized ERM by GF/GD has been studied extensively; our work lets us systematically use such results to fuel Bayesian inference.

We also note sampling at at low temperature is important in generative AI; the end of the reverse process in diffusion models is at low temperature. See e.g. Song and Ermon (2019); Song et al. (2021a,b); Song (2021). Moreover, in the temperature restriction $\beta = \Omega(d)$, we can replace d by the rank of $\nabla^2\Phi$ when Φ is smooth.

10. For example, it remains open for the Sherrington-Kirkpatrick model, for a PI/LSI w.r.t. the Glauber Dynamics.

3. Connecting Optimizability and Sampling

Before we state our results, we state the following unimodality assumption on F. Functional inequalities generally do not hold without exponential dimension-dependence when F has well-separated modes (Bovier et al., 2004, 2005; Menz and Schlichting, 2014). This is a probabilistic analogue to standard assumptions in non-convex optimization of good local behavior, such as F being convex or PŁ/KŁ near the global minima or near all saddle points, in e.g. Damian et al. (2021); Ahn et al. (2024).

Assumption 3.1 *Let $\mathcal{W}^\star$ denote the set of global minima. For all small enough $l > 0$, there exists $r(l) > 0$ such that $\{F \le l\} \subset \mathbb{B}(\mathcal{W}^\star, r(l))$ and $\mu_{\beta,\text{LOCAL}}(l)$, the restriction of μ_β on $\mathbb{B}(\mathcal{W}^\star, r(l))$, satisfies a Poincaré Inequality with constant $\mathsf{C}_{\text{PI, LOCAL}}(l)$. Here $\mathbb{B}(\mathcal{W}^\star, r(l)) = \{\mathbf{w} : d(\mathbf{w}, \mathcal{W}^\star) \le r(l)\}$, where $d(\cdot, \mathcal{W}^\star)$ denotes the distance from $\mathbf{w}$ to the closest point in $\mathcal{W}^\star$.*

Remark 10 *We believe the growth condition in the above is relatively unrestrictive. For example, if F is PŁ with parameter λ, by Theorem 2 of Karimi et al. (2016), $\{F \le l\} \subset \mathbb{B}(\mathcal{W}^\star, r(l))$ for $r(l) = 2\sqrt{l/\lambda}$.*[11]

Furthermore, there are several natural, general examples satisfying Assumption 3.1 subsuming standard settings of the literature with precise quantitative bounds on $\mathsf{C}_{\text{PI, LOCAL}}(l)$. We explain fully in Subsection A.3:

Example 1 *Suppose $\mathcal{W}^\star$ is convex and F is convex on $\mathbb{B}(\mathcal{W}^\star, r(l))$ for some $l > 0$. Then, we have that $\mathsf{C}_{\text{PI, LOCAL}}(l) \le \frac{(\text{diam}(\mathcal{W}^\star)+2r(l_b))^2}{\pi^2} = O(1)$ if $\text{diam}(\mathcal{W}) = O(1)$ (which is the case for $\beta = \Omega(d)$).*

Example 2 *Suppose also that F is α-strongly convex on $\mathbb{B}(\mathcal{W}^\star, r(l))$; then $\mathsf{C}_{\text{PI, LOCAL}}(l) = O\left(\frac{1}{\beta}\right)$. As a special case, consider the following stronger assumption in Lytras and Sabanis (2023) (see also Li and Erdogdu (2023)): $\mathcal{W}^\star = \{\mathbf{w}^\star\}$, F is α-strongly convex at $\mathbf{w}^\star$, and $\nabla^2 F$ is L'-Lipschitz in a $\Omega(1)$ neighborhood of $\mathbf{w}^\star$.*[12]

We emphasize that Assumption 3.1 or analogous assumptions are in fact *necessary*. For example, even $\mathcal{W}^\star$ being connected is not enough for efficient sampling.

Remark 11 *Consider when $\mathcal{W}^\star$ is dumbbell-shaped. Suppose $F(\mathbf{w}) = d(\mathbf{w}, \mathcal{W}^\star)^2$, where $d(\mathbf{w}, \mathbf{w}^\star)$ denotes the distance from $\mathbf{w}$ to the closest point in $\mathcal{W}^\star$. F is optimizable – its gradient is nonzero until reaching $\mathcal{W}^\star$.*[13] *However due to the poor isoperimetric constant of the dumbbell (Vempala, 2005), LMC will not mix rapidly upon reaching $\mathcal{W}^\star$, and so the isoperimetric constants of μ_β behave poorly for β large.*

To our knowledge, the only other related work handling multiple minimizers of F is Gong et al. (2024). Their result also deteriorates when $\mathcal{W}^\star$ has poor isoperimetric constant. Moreover, Assumption 3.1 does not directly imply a PI; terrible isoperimetry elsewhere gives poor mixing times from arbitrarily initialization. It also does not imply a WPI in terms of $\mathcal{S}$, the set where optimizability does not hold.

Convention. From here on out, asymptotic notation sometimes hides problem-dependent parameters; however we never suppress β, d-dependence. Explicit dependencies are written fully in the appendix.

3.1. Main Results: Poincaré and Log-Sobolev Inequalities

Consider the following assumption on the tail growth of F, which corresponds to linear tail growth of F, which goes hand-in-hand with a PI. We note only the second part of this assumption is required for smooth Φ.

11. See also Chewi and Stromme (2024), Otto and Villani (2000).
12. This applies for small enough l such that $\mathbb{B}(\mathcal{W}^\star, r(l))$ is a subset of this $\Omega(1)$ neighborhood.
13. One can straightforwardly check this verifies optimizability in our sense.

Assumption 3.2 *Suppose that for some $r_1, r_2, R > 0$, for all $\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R)^c$, we have $\langle \nabla F(\mathbf{w}), \mathbf{w} - \mathbf{w}^\star \rangle \geq r_1 F(\mathbf{w})$ and $F(\mathbf{w}) \geq r_2 \|\mathbf{w} - \mathbf{w}^\star\|$ for some $\mathbf{w}^\star \in \mathcal{W}^\star$.*

This assumption is very general in the context of optimization, and can be enforced via suitable regularization outside $\mathbb{B}(\mathbf{w}^\star, R)$ (Raginsky et al., 2017). The standard dissipativity assumption made in many prior works on non-convex optimization (Raginsky et al., 2017; Xu et al., 2018; Zou et al., 2021; Mou et al., 2022) are a special case of Assumption 3.2; consequently we present the assumption in the above form. Note as per Raginsky et al. (2017), for $\beta = \Omega(d)$, the dissipativity assumption (even dissipativity with $b = 0$) implies μ_β satisfies a PI, but with constant worst-case exponential in dimension.

Theorem 12 (Establishing PI and LSI under optimizability from all initializations) *Suppose F is optimizable in the sense of Definition 2 for all $\mathbf{w}$ and satisfies Assumption 3.2, the corresponding Φ satisfies Assumption 1.1 (F satisfying Assumption 1.1 is unnecessary here; see Remark 29), and Assumption 3.1 is satisfied for some $l_b > 0$. Then for $\beta \geq \Omega(d)$:*

1. *μ_β satisfies a PI with $\mathsf{C}_{\text{PI}} = O\Big(\mathsf{C}_{\text{PI, LOCAL}} + \frac{1}{\beta}\Big)$, where $\mathsf{C}_{\text{PI, LOCAL}}$ is the Poincaré constant of μ_β restricted to $\mathbb{B}(\mathcal{W}^\star, r(l_b))$.*
2. *Suppose F is L-weakly-convex, that is $\nabla^2 F(\mathbf{w}) \succeq -L\boldsymbol{I}_d$ for some $L > 0$, and F has quadratic tail growth, that is, $F(\mathbf{w}) \geq m\|\mathbf{w}\|^2 - b$ for some $m, b > 0$.[14] Let $S < \infty$ be the second moment of μ_β. Then μ_β satisfies a LSI with constant $\mathsf{C}_{\text{LSI}} = O\Big(\Big(S + \frac{d}{\beta} + 1\Big)(\beta \mathsf{C}_{\text{PI, LOCAL}} + 1)\Big)$.*

From Theorem 12, we have established that optimizability of F via GF/GD (under the conditions from above, among which Assumption 3.1 and Assumption 1.1 are needed) implies PI/LSI at low temperature. These inequalities are the *crux* of non-log-concave sampling via LMC. Central to this proof is the optimizability condition $\langle \nabla\Phi(\mathbf{w}), \nabla F(\mathbf{w}) \rangle \geq g(F(\mathbf{w}))$ from Definition 2; see Section C. As such, Theorem 12 confirms our initial Conjecture 1. Later in Subsection 3.3, we present corollaries of Theorem 12 for sampling.

Explicit constants are in the proof in Subsection D.1; they are not included for simplicity. To demonstrate one such example, consider when Φ is L-smooth, which as explained in Section 4 subsumes many cases of interest. Then we have the following, which we expand further on in Remark 30.

Remark 13 *Suppose Φ is L-smooth, $g(x) = \lambda x$ for $\lambda \leq 1$, and WLOG that $r_1 \leq 1/2$. Then μ_β satisfies a PI with*

$$\mathsf{C}_{\text{PI}} = 2\mathsf{C}_{\text{PI, LOCAL}} + \frac{2}{\beta}\left(1 + \frac{L}{\lambda l_b}\right) \text{ for } \beta \geq 2\left(1 + \frac{L}{\lambda l_b}\right)\left(d + \frac{8R^2}{r_1 L} \vee \frac{2L}{r_1 r_2^2 \lambda^2}\right). \tag{7}$$

Remark 14 *We also consider the sharpness of our guarantees. For $F(\mathbf{w}) = \|\mathbf{w}\|^2$ (and strongly convex F), the PI constant from Theorem 12 is tight up to $O(\cdot)$. The LSI constant is lossy by around a β factor.*

The proof of Theorem 12 uses the Lyapunov function technique in a fairly novel way. Typically one uses a particular ad-hoc Lyapunov function such as $e^{\beta F}$, F, or similar, as in Chewi and Stromme (2024); Gong et al. (2024); Lytras and Sabanis (2023); Li and Erdogdu (2023). Rather, we use Φ from Definition 2 – the *exact same* Lyapunov function arising from optimization (recall Definition 2, from De Sa et al. (2022)). We present the main ideas for the proof in Section C and the full proof in Section D.

3.2. Main Results: Weak Poincaré Inequalities

We now discuss how to extend our work to when optimizability in the form of Definition 2 holds in some region $\mathcal{S}$, where we prove a WPI. We establish the following; the proof is in Subsection D.2:

14. Recall quadratic tail growth goes hand-in-hand with a LSI.

Theorem 15 (Establishing WPI under optimizability from most initializations) *Suppose F is optimizable in the sense of Definition 2 for all $\mathbf{w}$ not in some $\mathcal{S} \subseteq \mathbb{R}^d$, F satisfies Assumption 3.2, F and the corresponding Φ satisfy Assumption 1.1, and Assumption 3.1 is satisfied for some $l_b > 0$. Then for all $\beta \geq \Omega(d)$, μ_β satisfies a $(\mathsf{C}_{\text{WPI}}, \delta)$-WPI with $\mathsf{C}_{\text{WPI}} = O\Big(\mathsf{C}_{\text{PI, LOCAL}} + \frac{1}{\beta}\Big)$, $\delta = O(\mu_\beta(\mathcal{S}))$.*

$\mathcal{S}$ typically has small Lebesgue measure ν, for example in the landscape of Phase Retrieval, Matrix Square Root, or the set of 'bad initializations' around a saddle point where Gradient Descent does not escape (Jain et al., 2017; Jin et al., 2017; Lee et al., 2019; De Sa et al., 2022). For $\beta \geq \Omega(d)$, $\mu_\beta(\mathcal{S}) \leq \frac{1}{Z}\exp(-\beta \inf_{\mathbf{w}\in\mathcal{S}} F(\mathbf{w}))\nu(\mathcal{S})$, where $Z = \int e^{-\beta F} \mathrm{d}\mathbf{w}$. Unless $\mathcal{S}$ already comprises of favorable near-global-optima or $\nu(\mathcal{S})$ is large, this term is small. A crude upper bound follows from Markov's Inequality. Moreover if F is L-smooth, for $\beta = \tilde{\Omega}(d)$, we can lower bound $Z \geq e^{-d\ln(\beta L/2\pi)}$; see 3.21, Raginsky et al. (2017).[15] Thus in this case, the term $\frac{1}{Z}\exp(-\beta \inf_{\mathbf{w}\in\mathcal{S}} F(\mathbf{w})) = e^{-\Omega(\beta)}$ is exponentially small.

Thus by (6), LMC can sample to accuracy $4\mu_\beta(\mathcal{S})\Big\|\frac{\mathrm{d}\pi_0}{\mathrm{d}\mu_\beta} - 1\Big\|_\infty^2 \leq \frac{1}{Z}\exp(-\beta\inf_{\mathbf{w}\in\mathcal{S}} F(\mathbf{w}))\nu(\mathcal{S})\Big\|\frac{\mathrm{d}\pi_0}{\mathrm{d}\mu_\beta} - 1\Big\|_\infty^2$. Thus if $\nu(\mathcal{S})$ is small and we have a warm start in that $\Big\|\frac{\mathrm{d}\pi_0}{\mathrm{d}\mu_\beta} - 1\Big\|_\infty^2$ is controlled, LMC can sample to high accuracy. This confirms the intuition in Conjecture 3.

Remark 16 *Suppose Φ, F are L-smooth, $g(x) = \lambda x$ for $\lambda \leq 1$, and WLOG that $r_1 \leq 1/2$. Then μ_β satisfies a*

$$\left(2\mathsf{C}_{\text{PI, LOCAL}} + \frac{2}{\beta}\left(1 + \frac{B}{\lambda l_b}\right), 6\left(1 + \frac{B}{\lambda l_b}\right)\mu_\beta(\mathcal{S})\right) - \textit{WPI for } \beta \geq 2\left(1 + \frac{B}{\lambda l_b}\right)(d + C''),$$

where $B = L \vee G_F G_\Phi \vee 1$, $G_F = \sup_{\mathbf{w}\in\mathcal{S}\cap\mathbb{B}(\mathbf{w}^\star, R+1)}\|\nabla F(\mathbf{w})\|$, $G_\Phi = \sup_{\mathbf{w}\in\mathcal{S}\cap\mathbb{B}(\mathbf{w}^\star, R+1)}\|\nabla\Phi(\mathbf{w})\|$, $C'' = (\lambda+1)\left(\frac{8R^2}{r_1 L} \vee \frac{2L}{r_1 r_2^2 \lambda^2}\right) + \lambda G_F^2$. Notice in $\mathcal{S}$, the region where GF/GD do not succeed, we except G_F to be very small; if $\Phi = F$ (e.g. for PŁ, KŁ functions), we also obtain that G_Φ is small.

Corollary 17 (Of the proof; relaxing Assumption 3.1) *Suppose $\mu_{\beta,\text{LOCAL}}$ satisfies a $(\mathsf{C}_{\text{WPI, LOCAL}}, \delta_{\text{LOCAL}})$-WPI. Then in the setting of Theorem 12, μ_β satisfies a $\Big(O(\mathsf{C}_{\text{WPI, LOCAL}} + \frac{1}{\beta}), 2\delta_{\text{LOCAL}}\Big)$-WPI. Analogously in the setting of Theorem 15, μ_β satisfies a $\Big(O(\mathsf{C}_{\text{WPI, LOCAL}} + \frac{1}{\beta}), O\big(\mu_\beta(\mathcal{S}) + 2\delta_{\text{LOCAL}}\big)\Big)$-WPI.*

Remark 18 (Sampling with only Local Optimizability) *We further note that upon examining the proofs of Theorem 12, Theorem 15, we only need Definition 2 within $\mathbb{B}(\mathbf{w}^\star, R+1)$ for some $\mathbf{w}^\star \in \mathcal{W}^\star$. This suggests the following interesting algorithmic implication: if Definition 2 only holds locally in $\mathbb{B}(\mathbf{w}^\star, R)$, with advance knowledge of $\mathbf{w}^\star$ and R, one can still approximately sample from μ_β by regularizing F so Assumption 3.2 holds. We elaborate further in Subsection A.2; in particular see Proposition 21, Corollary 22.*

3.3. Algorithmic Implications for Sampling

We now state direct algorithmic implications of Theorem 12, Theorem 15. We remark Theorem 15 yields sampling results for the Langevin Diffusion (1) under a suitable warm start, via (6) (from Theorem 2.1, Röckner and Wang (2001)). Now we will focus on the implications of Theorem 12. Note establishing improved sampling algorithms under isoperimetry is *not* the main focus of our work; the following results are rather *corollaries* of Theorem 12 via the literature. Again, we believe this is a core *strength* of our work; our results *complement* the literature. Note several recent works have shown the success of discrete-time LMC under solely a PI and smoothness in TV and KL divergences, e.g. Chewi et al. (2024); Chen et al. (2022); Altschuler and Chewi (2024). We now are in position to state these implications.

Assumption 3.3 (L-Hölder-smoothness) *For any $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$, $\|\nabla F(\mathbf{w}_1) - \nabla F(\mathbf{w}_2)\| \leq L\|\mathbf{w}_1 - \mathbf{w}_2\|^s$.*

15. This requires an additional polylog factors.

Corollary 19 *Suppose F is optimizable by GF in the sense of Definition 2, the other conditions of Theorem 12 hold, and F satisfies Assumption 3.3. Then for all $\beta \geq \Omega(d)$, where the $\Omega(\cdot)$ hides F-dependent parameters, discrete-time LMC initialized at a distribution $\pi_0 \sim \mathcal{N}\Big(\vec{\mathbf{0}}, \frac{1}{2\beta L+\gamma}\boldsymbol{I}_d\Big)$ with appropriate step size has the following guarantees, where $\gamma \leq 1$ is defined in our proof in Subsection D.3.*

1. *Suppose F satisfies Assumption 3.3, that is, F is L-Hölder-continuous with parameter s in $(0,1]$. Then with access to a gradient oracle ∇F, the recursion (2) yields a distribution π_T with $\mathsf{TV}\big(\pi_T \| \mu_\beta\big) \leq \varepsilon$* after $T = \tilde{O}\Big(d(\mathsf{C}_{\text{PI, LOCAL}} + \frac{1}{\beta})^{1+\frac{1}{s}}\beta^{1+\frac{3}{s}} \max\Big\{1, \frac{\beta^{s/2}}{d}\Big\}\varepsilon^{-\frac{2}{s}}\Big)$ iterations.
2. *Suppose that F is L-smooth. Given additional access to a Proximal Oracle, the Proximal Sampler yields μ_T with $\mathsf{d}\big(\pi_T \| \mu_\beta\big) \leq \varepsilon$* after $T = \tilde{O}\Big(\big(\mathsf{C}_{\text{PI, LOCAL}} + \frac{1}{\beta}\big)\beta d^{1/2}\big\{\beta + d + \log\big(\frac{1}{\varepsilon}\big)\big\}\Big)$ iterations, *in the metrics $\mathsf{d} \in \{\mathsf{TV}, \sqrt{\mathsf{KL}}, \sqrt{\chi^2}\}$. See Subsection B.2 for more details on the Proximal Sampler.*

We discuss further details on how the above follows from the literature in Subsection D.3. Furthermore, note Assumption 3.3 does not capture many (optimizable) F of interest, for example simply $F(x) = x^{2p}$ for any $p \geq 1$ in one dimension. In Subsection A.1 we discuss how we can adapt the recent work Lytras and Mertikopoulos (2024) to such situations; see Corollary 20. Note in both of Corollary 19, Corollary 20, we do not use information about $\mathcal{W}^\star$ in the initialization, and do not make a warm start hypothesis. The initialization $\frac{1}{\beta}\boldsymbol{I}_d$ is for similar scaling as μ_β, needed to control initialization, and $\vec{\mathbf{0}}$ is arbitrary.[16] Our sampling algorithms do not use knowledge of F in initialization; they succeed because the success of GF/GD imply isoperimetry, as per Theorem 12. Intuitively, the optimizability condition $\langle \nabla\Phi(\mathbf{w}), \nabla F(\mathbf{w})\rangle \geq g(F(\mathbf{w}))$ allows gradient-based LMC to 'find' $\mathcal{W}^\star$ without a warm start.

4. Examples and Applications

The framework of 'optimizability' from Definition 2 and Assumption 1.1 subsumes many interesting examples in non-convex (and convex) optimization, from smooth PŁ and KŁ functions to Phase Retrieval and Matrix Square Root to *all* Linearizable functions; see De Sa et al. (2022). In all these examples (3) holds, and Assumption 1.1 is satisfied with *dimension-independent* ρ_Φ. Combining with the conditions of Theorem 12, Corollary 19, Corollary 20, we obtain results on isoperimetry and sampling via LMC for many examples.

Example 3 (PŁ functions) *Consider smooth PŁ functions F, that is with $\|\nabla F(\mathbf{w})\|^2 \geq \lambda F(\mathbf{w})$. Then Definition 2 holds with $\Phi(\mathbf{w}) = F(\mathbf{w})$ and $g(x) = \lambda x$. Note Assumption 1.1 holds as F is smooth. Note also that F need not be smooth; we only need Assumption 1.1 to hold with F in place of Φ. For example, for $\rho_\Phi(x) = A'x + B'$, Assumption 1.1 allows for arbitrary polynomial tail growth of F in $\|\mathbf{w}\|$.*

Example 4 (KŁ functions) *Now consider KŁ functions F, that is with $\|\nabla F(\mathbf{w})\|^2 \geq \lambda F(\mathbf{w})^{1+\theta}$ for $\theta \geq 0$. The main difference between the PŁ and KŁ conditions is that the KŁ condition is weaker near the global minima. For KŁ functions F, we can take $\Phi(\mathbf{w}) = \frac{F(\mathbf{w})}{\lambda}$ in the above, and Definition 2 holds with $g(x) = x^{1+\theta}$, if F satisfies Assumption 1.1 with Φ in place of F. Again, note Assumption 1.1 holds if F is smooth by the definition of smoothness and Lemma 42, but that F satisfying Assumption 1.1 is much more general than F being smooth, and in particular allows for any polynomial tail growth of F in $\|\mathbf{w}\|$.*

Example 5 (Linearizable/Quasar-Convex Functions) *Consider λ-Linearizable functions F (Kale et al., 2021), that is s.t. $\langle \nabla F(\mathbf{w}), \mathbf{w} - \mathbf{w}^\star\rangle \geq \lambda F(\mathbf{w})$ (also known as Quasar-Convexity, see Definition 3 of Hinder et al. (2020), or Weak Quasi-Convexity, see Hardt et al. (2018)). Here $\Phi(\mathbf{w}) = \|\mathbf{w} - \mathbf{w}^\star\|^2$ and $g(x) = \lambda x$, and Definition 2 holds. Note Φ, being 2-smooth, vacuously satisfies Assumption 1.1 by Lemma 42. For a PI (Theorem 12), Assumption 1.1 is not needed on F, and thus we obtain a PI with no regularity assumptions on*

16. The initial divergence can be controlled in Lemma 43, Lemma 46, and these divergences already factor into our runtime bounds.

F. One can obtain the β-range for which one obtains a PI from our results by taking $L = 2$ in (7). This setting generalizes numerous other non-convex function classes from optimization, such as star-convex functions (Lee and Valiant, 2016) and smooth one-point-strongly convex functions (Kleinberg et al., 2018). See Hinder et al. (2020) for further discussion.

Applying our main results Theorem 12, Theorem 15, we obtain isoperimetry for all these examples (under the conditions of those Theorems). Noting Assumption 3.1 is satisfied automatically for all convex F, combining Theorem 12 with Corollary 20 gives sampling results for log-concave measures beyond smoothness. Formal statements of these corollaries are in Corollary 23, Corollary 24.

5. High-Level Proof Ideas

Here, we give high-level proof ideas. A more detailed proof sketch and full proofs are in Section C and Section D respectively. The central idea is to prove a PI via the Lyapunov potential Φ from optimization using the Lyapunov function technique from Bakry et al. (2008). We develop the technique to fully exploit the property (3) implied by success of GF/GD, which gives us sharp quantitative control of the isoperimetric constant. For simplicity, we suppose here that Φ is L-smooth and that $g(x) = x$.

Proving a PI: Let $\mathcal{U} = \mathbb{B}(\mathcal{W}^\star, r(l_b))$, where l_b is any l satisfying Assumption 3.1. Consider an arbitrary test function ψ. Let $f = \psi - \alpha$, where $\alpha = \int_{\mathcal{U}} \psi \mathrm{d}\mu_{\beta,\text{LOCAL}}$. For $B > 0$ to be chosen later, note as $\frac{t}{t+B}$ is increasing in $t \geq 0$, we have

$$\frac{l_b}{l_b + B}\mathbb{V}_{\mu_\beta}[\psi] \leq \frac{l_b}{l_b + B}\int f^2 \mathrm{d}\mu_\beta \leq \frac{l_b}{l_b + B}\int_{\mathcal{U}} f^2 \mathrm{d}\mu_\beta + \frac{l_b}{l_b + B}\int_{\mathcal{U}^c} f^2 \mathrm{d}\mu_\beta$$
$$\leq \frac{l_b}{l_b + B}\int_{\mathcal{U}} f^2 \mathrm{d}\mu_\beta + \int f^2 \frac{F(\mathbf{w})}{F(\mathbf{w}) + B}\mathrm{d}\mu_\beta.$$

We upper bound the first integral $\int_{\mathcal{U}} f^2 \mathrm{d}\mu_\beta$ by Assumption 3.1 and the choice of α. For the second integral, note by the condition (3) and letting $\mathcal{L}$ denote the so-called *infinitesimal generator* of (1), we have

$$F(\mathbf{w}) + B \leq \langle \nabla\Phi(\mathbf{w}), F(\mathbf{w})\rangle + B = -\frac{1}{\beta}\mathcal{L}\Phi(\mathbf{w}) + \frac{1}{\beta}\Delta\Phi(\mathbf{w}) + B \leq -\frac{1}{\beta}\mathcal{L}\Phi(\mathbf{w}) + \frac{1}{\beta}|\Delta\Phi(\mathbf{w})| + B.$$

We divide by $F(\mathbf{w}) + B > 0$, multiply both sides by $f^2 \geq 0$, and integrate with respect to μ_β to obtain

$$\int f^2 \frac{F(\mathbf{w})}{F(\mathbf{w}) + B}\mathrm{d}\mu_\beta \leq \frac{1}{\beta}\int f(\mathbf{w})^2 \frac{-\mathcal{L}\Phi(\mathbf{w})}{F(\mathbf{w}) + B}\mathrm{d}\mu_\beta + \frac{1}{\beta}\int f(\mathbf{w})^2 \frac{|\Delta\Phi(\mathbf{w})|}{F(\mathbf{w}) + B}\mathrm{d}\mu_\beta.$$

We upper bound the first integral above using properties of the infinitesimal generator, Lemma 26. We upper bound the second integral above using smoothness of Φ and that $\beta = \Omega(d)$. Rearranging and converting back to ψ gives the desired PI. To generalize the proof to non-smooth Φ, we 'interpolate' Φ with the smooth function $\|\mathbf{w} - \mathbf{w}^\star\|^2$ and use Assumption 3.2.

Proving a WPI: We follow the same steps as above, except we apply the above inequality *pointwise*, for $\mathbf{w} \in \mathcal{S}^c$ where it holds. We use this to upper bound $\int f^2 \mathrm{d}\mu_\beta$ in a similar fashion as above, which in turn lets us upper bound $\int f^2 \frac{F(\mathbf{w})}{F(\mathbf{w})+B}\mathrm{d}\mu_\beta$. The difference is that we pick up an 'error term' $\int_{\mathcal{S}} f^2 \mathrm{d}\mu_\beta$. However by definition of f, we have $f^2 \leq \mathrm{osc}(\psi)^2$, and so the error term is at most $\mathrm{osc}(\psi)^2 \mu_\beta(\mathcal{S})$.

Acknowledgments

We would like to thank Ayush Sekhari for collaboration on initial stages of the project and generously reading an earlier manuscript, Ahmed El Alaoui for a related collaboration which inspired part of our work, and Brice Huang and Robert Kleinberg for discussions.

References

Kwangjun Ahn, Ali Jadbabaie, and Suvrit Sra. How to Escape Sharp Minima with Random Perturbations. *Forty-first International Conference on Machine Learning*, 2024.

Ahmed El Alaoui, Ronen Eldan, Reza Gheissari, and Arianna Piana. Fast relaxation of the random field Ising dynamics. *Annals of Probability (to appear)*, 2025+.

Jason M Altschuler and Sinho Chewi. Faster high-accuracy log-concave sampling via algorithmic warm starts. *Journal of the ACM*, 71(3):1–55, 2024.

Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84: Proceedings*, pages 177–206. Springer, 2006.

Dominique Bakry, Franck Barthe, Patrick Cattiaux, and Arnaud Guillin. A simple proof of the Poincaré inequality for a large class of probability measures. *Electronic Communications in Probability*, 13:60–66, 2008.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. In *Conference on Learning Theory*, 2018.

Sergey G Bobkov. Isoperimetric and analytic inequalities for log-concave probability measures. *The Annals of Probability*, 27(4):1903–1921, 1999.

Michel Bonnefont. Poincaré inequality With Explicit Constant In Dimension $d \geq 1$. *https://www.math.u-bordeaux.fr/ mibonnef/Poincaré__Toulouse.pdf*, 2022.

Dominique Bontemps. Bernstein–von mises theorems for Gaussian regression with increasing number of regressors. *Annals of Statistics*, 39, 2011.

Anton Bovier, Michael Eckhoff, Véronique Gayrard, and Markus Klein. Metastability in reversible diffusion processes. I. Sharp asymptotics for capacities and exit times. *Journal of the European Mathematical Society*, 6(4):399–424, 2004.

Anton Bovier, Véronique Gayrard, and Markus Klein. Metastability in reversible diffusion processes II: Precise asymptotics for small eigenvalues. *Journal of the European Mathematical Society*, 7(1):69–99, 2005.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University press, 2004.

Herm Jan Brascamp and Elliott 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, 1976.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

Patrick Cattiaux, Arnaud Guillin, and Li-Ming Wu. A note on Talagrand's transportation inequality and logarithmic Sobolev inequality. *Probability Theory and Related Fields*, 148:285–304, 2010.

August Y Chen, Ayush Sekhari, and Karthik Sridharan. Langevin Dynamics: A Unified Perspective on Optimization via Lyapunov Potentials. *arXiv preprint arXiv:2407.04264*, 2024.

Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono. Improved analysis for a proximal algorithm for sampling. In *Conference on Learning Theory*, pages 2984–3014. PMLR, 2022.

Sinho Chewi. Log-concave sampling. *Book draft available at https://chewisinho. github.io*, 2024.

Sinho Chewi and Austin J Stromme. The ballistic limit of the log-Sobolev constant equals the Polyak-Lojasiewicz constant. *arXiv preprint arXiv:2411.11415*, 2024.

Sinho Chewi, Murat A Erdogdu, Mufan Li, Ruoqi Shen, and Matthew S Zhang. Analysis of Langevin Monte Carlo from Poincaré to Log-Sobolev. *Foundations of Computational Mathematics*, pages 1–51, 2024.

Tzuu-Shuh Chiang, Chii-Ruey Hwang, and Shuenn Jyi Sheu. Diffusion for global optimization in r̂n. *SIAM Journal on Control and Optimization*, 25(3):737–753, 1987.

Alex Damian, Tengyu Ma, and Jason D Lee. Label Noise SGD Provably Prefers Flat Global Minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461, 2021.

Aniket Das, Dheeraj M Nagaraj, and Anant Raj. Utilising the CLT Structure in Stochastic Gradient based Sampling: Improved Analysis and Faster Algorithms. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4072–4129. PMLR, 2023.

Christopher M De Sa, Satyen Kale, Jason D Lee, Ayush Sekhari, and Karthik Sridharan. From Gradient Flow on Population Loss to Learning with Stochastic Gradient Descent. *Advances in Neural Information Processing Systems*, 35:30963–30976, 2022.

Ahmed El Alaoui and Jason Gaitonde. Bounds on the covariance matrix of the Sherrington–Kirkpatrick model. *Electronic Communications in Probability*, 29:1–13, 2024.

Jiaojiao Fan, Bo Yuan, and Yongxin Chen. Improved dimension dependence of a proximal algorithm for sampling. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1473–1521. PMLR, 2023.

Dylan J Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. *Advances in Neural Information Processing Systems*, 2018.

Dani Gamerman and Hedibert F Lopes. *Markov Chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC, 2006.

Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov Chain Monte Carlo in practice*. CRC press, 1995.

Yun Gong, Niao He, and Zebang Shen. Poincaré Inequality for Local Log-Polyak-Lojasiewicz Measures: Non-asymptotic Analysis in Low-temperature Regime. *arXiv preprint arXiv:2501.00429*, 2024.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018.

Oliver Hinder, Aaron Sidford, and Nimit Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. In *Conference on Learning Theory*, pages 1894–1938. PMLR, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Brice Huang and Mark Sellke. Strong topological trivialization of multi-species spherical spin glasses. *Annals of Probability (to appear)*, 2025+.

Brice Huang, Sidhanth Mohanty, Amit Rajaraman, and David X Wu. Weak Poincaré inequalities, Simulated Annealing, and Sampling from Spherical Spin Glasses. In *Symposium on Theory of Computation*, 2025.

Xunpeng Huang, Difan Zou, Hanze Dong, Yian Ma, and Tong Zhang. Faster Sampling via Stochastic Gradient Proximal Sampler. *Forty-first International Conference on Machine Learning*, 2024.

Prateek Jain, Chi Jin, Sham Kakade, and Praneeth Netrapalli. Global convergence of non-convex gradient descent for computing matrix squareroot. In *Artificial Intelligence and Statistics*, pages 479–488. PMLR, 2017.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to Escape Saddle Points Efficiently. In *International Conference on Machine Learning*, pages 1724–1732. PMLR, 2017.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

Satyen Kale, Ayush Sekhari, and Karthik Sridharan. SGD: The Role of Implicit Regularization, Batch-size and Multiple-epochs. *Advances In Neural Information Processing Systems*, 34:27422–27433, 2021.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pages 795–811. Springer, 2016.

Yuri Kinoshita and Taiji Suzuki. Improved convergence rate of stochastic gradient Langevin dynamics with variance reduction and its application to optimization. *Advances in Neural Information Processing Systems*, 35:19022–19034, 2022.

Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does SGD escape local minima? In *International Conference on Machine Learning*, pages 2698–2707. PMLR, 2018.

Yunbum Kook and Santosh S Vempala. Gaussian Cooling and Dikin Walks: The Interior-Point Method for Logconcave Sampling. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 3137–3240. PMLR, 2024.

Dirk P Kroese, Thomas Taimre, and Zdravko I Botev. *Handbook of monte carlo methods*. John Wiley & Sons, 2013.

Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier*, 48(3):769–783, 1998.

Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, 176:311–337, 2019.

Jasper CH Lee and Paul Valiant. Optimizing star-convex functions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science*, pages 603–614. IEEE, 2016.

Yin Tat Lee, Ruoqi Shen, and Kevin Tian. Structured logconcave sampling with a restricted Gaussian oracle. In *Conference on Learning Theory*, pages 2993–3050. PMLR, 2021.

Joseph Lehec. The Langevin Monte Carlo algorithm in the non-smooth log-concave case. *The Annals of Applied Probability*, 33(6A):4858–4874, 2023.

Mufan Li and Murat A Erdogdu. Riemannian langevin algorithm for solving semidefinite programs. *Bernoulli*, 29(4):3093–3113, 2023.

Jiaming Liang and Yongxin Chen. A proximal algorithm for sampling. *arXiv preprint arXiv:2202.13975*, 2022a.

Jiaming Liang and Yongxin Chen. A proximal algorithm for sampling from non-smooth potentials. In *2022 Winter Simulation Conference*, pages 3229–3240. IEEE, 2022b.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

Stanislaw Lojasiewicz. A topological property of real analytic subsets. *Coll. du CNRS, Les équations aux dérivées partielles*, 117(87-89):2, 1963.

Iosif Lytras and Panayotis Mertikopoulos. Tamed Langevin sampling under weaker conditions. *arXiv preprint arXiv:2405.17693*, 2024.

Iosif Lytras and Sotirios Sabanis. Taming under isoperimetry. *arXiv preprint arXiv:2311.09003*, 2023.

Yi-An Ma, Yuansi Chen, Chi Jin, Nicolas Flammarion, and Michael I Jordan. Sampling can be faster than optimization. *Proceedings of the National Academy of Sciences*, 116(42):20881–20885, 2019.

Georg Menz and André Schlichting. Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape. *Annals of Probability*, 42(5):1809–1884, 2014.

Andrea Montanari and Yuchen Wu. Posterior Sampling in High Dimension via Diffusion Processes. *arXiv:2304.11449*, 2023.

Wenlong Mou, Nicolas Flammarion, Martin J Wainwright, and Peter L Bartlett. Improved bounds for discretization of Langevin diffusions: Near-optimal rates without convexity. *Bernoulli*, 28(3):1577–1601, 2022.

Alireza Mousavi-Hosseini, Tyler K Farghly, Ye He, Krishna Balasubramanian, and Murat A Erdogdu. Towards a complete analysis of Langevin Monte Carlo: Beyond Poincaré Inequality. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1–35. PMLR, 2023.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

Lawrence E Payne and Hans F Weinberger. An optimal Poincaré inequality for convex domains. *Archive for Rational Mechanics and Analysis*, 5(1):286–292, 1960.

Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.

Michael Röckner and Feng-Yu Wang. Weak Poincaré inequalities and L2-convergence rates of Markov semigroups. *Journal of Functional Analysis*, 185(2):564–603, 2001.

Mark Sellke. The threshold energy of low temperature Langevin dynamics for pure spherical spin glasses. *Communications on Pure and Applied Mathematics*, 77(11):4065–4099, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. *International Conference on Learning Representations*, 2021a.

Yang Song. Generative Modeling by Estimating Gradients of the Data Distribution. *https://yang-song.net/blog/2021/score/*, 2021.

Yang Song and Stefano Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021b.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. *Advances in Neural Information Processing Systems*, 23, 2010.

Andrew M Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

Michalis K Titsias and Omiros Papaspiliopoulos. Auxiliary gradient-based sampling algorithms. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):749–767, 2018.

Santosh Vempala. Geometric random walks: a survey. *Combinatorial and Computational Geometry*, 52 (573-612):2, 2005.

Santosh Vempala and Andre Wibisono. Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices. *Advances in Neural Information Processing Systems*, 32, 2019.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Society, 2021.

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

Fengyu Wang. *Functional Inequalities, Markov Semigroups and Spectral Theory*. Elsevier, 2006.

Jun-Kun Wang and Andre Wibisono. Continuized Acceleration for Quasar-Convex Functions in Non-Convex Optimization. *International Conference on Learning Representations*, 2023.

Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global Convergence of Langevin Dynamics Based Algorithms for Nonconvex Optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *International Conference on Learning Representations*, 2020.

Difan Zou, Pan Xu, and Quanquan Gu. Faster Convergence of Stochastic Gradient Langevin Dynamics for Non-Log-Concave Sampling. In *Uncertainty in Artificial Intelligence*, pages 1152–1162. PMLR, 2021.

Contents

Notation. The domain is $\mathbb{R}^d$, with origin $\vec{\mathbf{0}}$. Let ν denote Lebesgue measure on $\mathbb{R}^d$. When we write $\|\cdot\|$ without explicitly specifying, we mean the l_2 Euclidean norm of a vector. For vectors $\mathbf{a}, \mathbf{b}$, let $\theta\langle \vec{a}, \vec{b} \rangle$ denote the directed angle they make in $[0, \pi]$. We denote the Laplacian (sum of second derivatives) of a twice-differentiable function f by Δf. We denote the Euclidean l_2 ball centered at $p \in \mathbb{R}^d$ with radius $R \geq 0$ by $\mathbb{B}(p, R)$. When $\mathcal{P}$ is a set, $\mathbb{B}(\mathcal{P}, R) = \{\mathbf{w} : \inf_{\mathbf{w}' \in \mathcal{P}} \|\mathbf{w} - \mathbf{w}'\| \leq R\}$. We denote the surface of the d-dimensional sphere with radius r by $\mathcal{S}^{d-1}(r)$. For some f differentiable to k orders, we will let $\nabla^k f$ denote the tensor of all the k-th order derivatives of f, and $\|\cdot\|_{\mathrm{op}}$ denotes the corresponding tensor's operator norm. For a matrix $\boldsymbol{M}$, let $\lambda_{\min}(\boldsymbol{M})$ denote its minimum eigenvalue, and $\mathrm{tr}(\boldsymbol{M})$ denote its trace. For matrices

$\boldsymbol{M}_1, \boldsymbol{M}_2$, we let $\succeq$ denote the PSD order, that is $\boldsymbol{M}_1 \succeq \boldsymbol{M}_2$ if and only if $\boldsymbol{M}_1 - \boldsymbol{M}_2$ is positive semi-definite. We denote Total Variation distance, Kullback–Leibler divergence, and Chi-squared divergence by TV, KL, χ^2 respectively.

For an arbitrary function f, let $\operatorname{osc}(f) = \sup f - \inf f$. Here, $\widetilde{\Omega}, \widetilde{\Theta}, \widetilde{O}$ hide universal constants and $\log$ factors in β, d, ε. We denote the set of all global minimizers $\mathbf{w}^\star$ of F by $\mathcal{W}^\star$. We say F is *smooth* (L-smooth) if the magnitude of the eigenvalues of its Hessian are universally bounded by a constant (when this constant is at most L). We let Z denote the partition function of the corresponding measure, which may change line-by-line (e.g. for different β).

Appendix A. Additional Results and Discussion

A.1. Further Algorithmic Implications of Main Results

The assumption of smoothness or Hölder continuity does not capture many optimizable F of interest, for example simply $F(x) = x^{2p}$ for any $p > 1$ in one dimension. See e.g. Zhang et al. (2020) and follow-ups for a study of optimizable F which are not smooth. We thus consider a more general assumption from Lytras and Mertikopoulos (2024) (their Assumption 1, slightly simplified) which goes far beyond, allowing for tail growth of F that is any arbitrary polynomial in $\|\mathbf{w}\|$. In particular, this assumption can be verified for $F(x) = x^{2p}$, which is not true for Assumption 3.3. Under this assumption, we obtain less sharp, but still polynomial, convergence rates:

Assumption A.1 (Almost Assumption 1, Lytras and Mertikopoulos (2024)) *F satisfies the following:*

- *Weak Dissipativity: for some $s_2 \geq 1$, $A_2, b_2 > 0$, we have for all $\mathbf{w} \in \mathbb{R}^d$, $\langle \nabla F(\mathbf{w}), \mathbf{w} \rangle \geq A_2 \|\mathbf{w}\|^{s_2} - b_2$.*
- *Polynomial Jacobian Growth: for some $L_3, s_3 > 0$ and all $k \geq 2$ for which the following is well-defined, we have for all $\mathbf{w} \in \mathbb{R}^d$, $\max\left(\|\nabla F(\mathbf{w})\|, \left\|\nabla^k F(\mathbf{w})\right\|_{\mathrm{op}}\right) \leq L_3(1 + \|\mathbf{w}\|)^{2s_3}$.*

We emphasize we do *not* use these assumptions to obtain isoperimetry in Theorem 12. Rather, they are just different regularity assumptions under which we obtain different rates for discrete-time LMC. Under these assumptions, and recalling all dependence on d, β is polynomial in Theorem 12, we obtain from Theorem 12 that:

Corollary 20 *Suppose the conditions of Theorem 12 hold and F satisfies Assumption A.1. Moreover, suppose we initialize at a distribution $\pi_0 \propto \exp\left(-2\|\mathbf{w}\|^{2s_3'}\right)$ where $s_3' = \max(s_3 + \frac{1}{2}, r + 1)$, $r \geq \max(2s_3 + 1, s_3 + 2)$. Then assuming knowledge of A_2, s_1, s_2, s_3 from Assumption A.1 and with this initialization π_0, for $\beta = \Omega(d)$, discrete-time LMC enjoys the following guarantees:*

1. *Via the discrete-time algorithm Regularized Tamed Unadjusted Langevin (reg-TULA) of Lytras and Mertikopoulos (2024), we have $\mathsf{TV}\left(\pi_T \| \mu_\beta\right) \leq \varepsilon$ after $T = \tilde{O}\left(\mathrm{poly}\left(d, \beta, \mathsf{C}_{\text{PI, LOCAL}}, \frac{1}{\varepsilon}\right) \log\left(\frac{1}{\varepsilon}\right)\right)$ iterations.*

2. *Suppose the assumptions in point 2 of Theorem 12 also hold, which implies μ_β satisfies a Log-Sobolev Inequality with constant $\mathsf{C}_{\text{LSI}} = O(\beta \max\{S, 1\} \max\{\mathsf{C}_{\text{PI, LOCAL}}, 1\})$. Then via the discrete-time algorithm Weakly Dissipative Tamed Unadjusted Langevin Algorithm (wd-TULA) of Lytras and Mertikopoulos (2024), we have $\mathsf{TV}\left(\pi_T \| \mu_\beta\right) \leq \varepsilon$ after $T = \tilde{O}\left(\frac{poly(d,\beta) \max\{S,1\} \max\{\mathsf{C}_{\text{PI, LOCAL}},1\}}{\varepsilon^2} \log\left(\frac{1}{\varepsilon}\right)\right)$ iterations.*

Both of these sampling algorithms from Lytras and Mertikopoulos (2024) are fully detailed in Subsection B.3. Explicit polynomial dependencies can be found in the proof of Theorems 2, 3 from Lytras and Mertikopoulos (2024); the degrees of these polynomials depend (polynomially) on s_2, s_3.

A.2. Sampling Under Local Optimizability

Suppose rather than global optimizability, F is optimizable by GF only in a large region around $\mathbf{w}^\star$. Such a situation has been often observed in non-convex landscapes, for example in neural networks (Kleinberg et al., 2018; Liu et al., 2022). Rather than a WPI, we aim to prove a PI/LSI here for a regularized version of μ_β, and discuss its algorithmic implications. We impose the following regularity assumption on F:

Assumption A.2 *F is L-smooth for all $\mathbf{w}$, and for some $R > 0$:*

- *F is optimizable in $\mathbb{B}(\mathbf{w}^\star, R)$ where g in (3) is of the form $g(x) = \lambda x$ for $\lambda \le 1$.*
- *$\langle \nabla F(\mathbf{w}), \mathbf{w} - \mathbf{w}^\star \rangle \ge 0$ for all $\mathbf{w}$ with $R - 1 \le \|\mathbf{w} - \mathbf{w}^\star\| \le R$.*
- *$F(\mathbf{w}) \ge r_2 \|\mathbf{w} - \mathbf{w}^\star\|$ for some $r_2 > 0$.*

We can replace the smoothness assumption with Assumption A.1 by changing the regularization added to F appropriately, and can also replace 1 in the second condition $R - 1 \le \|\mathbf{w} - \mathbf{w}^\star\| \le R$ above by an arbitrary universal constant; see the proof in Subsection D.4. The condition on $g(\cdot)$ above is made for simplicity, and already captures several examples, e.g. PŁ and Linearizable functions; again, by suitably modifying the proof one can extend this to general $g(\cdot)$ satisfying the conditions of Definition 2. We stick with the above and discuss in Remark 40 how to generalize the proof.

Note here that outside $\mathbb{B}(\mathbf{w}^\star, R)$, besides smoothness and a lower bound on growth, F could have arbitrarily many points with vanishing gradient, saddle points and local minima.[17] This contrasts to the main result of Gong et al. (2024), where despite its supposed 'local' nature, their Assumption 5 lower bounds on $\|\nabla F\|$ and the lack of saddle points are assumed outside a compact set.

By regularizing F appropriately, we establish:

Proposition 21 *Suppose Assumption A.2 holds, the corresponding Φ satisfies Assumption 1.1, and Assumption 3.1 is satisfied for some $l_b > 0$ with $\mathbb{B}(\mathcal{W}^\star, r(l_b)) \subseteq \mathbb{B}(\mathbf{w}^\star, R - 1)$ for some $\mathbf{w}^\star \in \mathcal{W}^\star$. Let $\hat{F}(\mathbf{w}) = F(\mathbf{w}) + \chi_F(\mathbf{w}) \cdot L(\|\mathbf{w} - \mathbf{w}^\star\|^2 + 1)$ where $\chi_F \in [0, 1]$ is a suitable interpolant which depends on problem parameters, defined in our proof in (41).*

Then for $\beta \ge \Omega(d)$, $\hat{\mu}_\beta \propto e^{-\hat{F}}/Z$ satisfies a PI with constant $O\left(\mathsf{C}_{\text{PI, LOCAL}} + \frac{1}{\beta}\right)$. Furthermore, $\hat{F}$ is smooth with $O(1)$ smoothness constant.

Explicit constants are in the proof in Subsection D.4. We note that under the conditions of point 2 of Theorem 12 and via the same proof, one can also extend this to an LSI. As an algorithmic implication, Proposition 21 directly shows the following.

Corollary 22 *Let $\delta = \mu_\beta(\mathbb{B}(\mathbf{w}^\star, R - 1)^c)$. Given oracle access to ∇F, F and knowledge of $\mathbf{w}^\star \in \mathcal{W}^\star$ satisfying the conditions of Proposition 21, R, and $g(\cdot)$, then running LMC in both continuous and discrete time with $\nabla \hat{F}$ in place of ∇F yields a distribution π such that $\mathsf{TV}(\pi, \mu_\beta) \le \varepsilon + 3\delta$, in time $O\left(\text{poly}(d, \beta, \frac{1}{\varepsilon})\right)$.*

Proof of Corollary 22. By Proposition 21 and Corollary 19, Corollary 20, in continuous and discrete time, LMC yields a distribution π such that $\mathsf{TV}(\pi, \hat{\mu}_\beta) \le \varepsilon$ in time $O\left(\text{poly}(d, \beta, \frac{1}{\varepsilon})\right)$. Note LMC is implementable because we can construct $\nabla \hat{F}$ using knowledge of ∇F, $\mathbf{w}^\star \in \mathcal{W}^\star$ satisfying the conditions of Proposition 21, R, and problem-dependent parameters. The problem-dependent parameters are defined in the proof of Subsection D.4, and can be computed with oracle access to $F, \nabla F$, knowledge of $\mathbf{w}^\star, R, g(\cdot)$, and appropriate cross validation; we expand on this in Remark 41 in Subsection D.4. Hence we can implement LMC and

17. Smoothness and the lower bound on growth do not 'sandwich' F in a way that implies a lack of critical points.

produce a hypothesis π which approximately samples from $\hat{\mu}_\beta$ as per the above. Thus, we have

$$\mathsf{TV}(\pi, \mu_\beta) \le \mathsf{TV}(\pi, \hat{\mu}_\beta) + \mathsf{TV}(\hat{\mu}_\beta, \mu_\beta) \le \varepsilon + 3\delta,$$

where the last step is verified in Lemma 36. ■

We conclude from Corollary 22 that optimizability from appropriate neighborhoods of the global minima yields sampling guarantees, via running LMC on a regularized version of F. Running LMC on a regularized version of F has seen recent interest, as a way to sample from μ_β under relaxed regularity assumptions (Lytras and Sabanis, 2023; Lytras and Mertikopoulos, 2024). Here we offer a novel perspective justifying the benefit of regularization for LMC as a way we can sample from a regularized Gibbs measure if we only have 'local optimizability', and possibly adversarial behavior outside of this neighborhood.

A.3. Further Discussion of Examples and Implications

We first expand on why the natural settings Example 1, Example 2 are subsumed by Assumption 3.1:

- Example 1: Suppose $\mathcal{W}^\star$ is convex and F is convex on $\mathbb{B}(\mathcal{W}^\star, r(l))$ for some $l > 0$. Note convexity of $\mathcal{W}^\star$ implies convexity of $\mathbb{B}(\mathcal{W}^\star, r(l))$ (Exercise 2.14, Boyd and Vandenberghe (2004)). By the Payne-Weinberger Theorem (Payne and Weinberger, 1960), in the form of Theorem 6.2 of Bonnefont (2022), we see $\mathsf{C}_{\text{PI, LOCAL}}(l) \le \frac{(\operatorname{diam}(\mathcal{W}^\star)+2r(l_b))^2}{\pi^2} = O(1)$ if $\operatorname{diam}(\mathcal{W}) = O(1)$. Note $\operatorname{diam}(\mathcal{W}) = O(1)$ is the case for $\beta = \Omega(d)$.
- Example 2: As a special case of the above, suppose additionally that F is α-strongly convex on $\mathbb{B}(\mathcal{W}^\star, r(l))$. Then $\mathsf{C}_{\text{PI, LOCAL}}(l) = O\left(\frac{1}{\beta}\right)$ by Brascamp-Lieb (Brascamp and Lieb, 1976) in the form of Theorem 5.1, Bonnefont (2022).[18] A special case of this is the following stronger assumption in Lytras and Sabanis (2023), also considered in Li and Erdogdu (2023): $\mathcal{W}^\star = \{\mathbf{w}^\star\}$, F is α-strongly convex at $\mathbf{w}^\star$, and the Hessian of F is L'-Lipschitz in a $\Omega(1)$ neighborhood of $\mathbf{w}^\star$. To see why, consider $l_b > 0$ small enough so that in $\mathbb{B}(\mathcal{W}^\star, r(l_b))$, the Hessian of F is L'-Lipschitz, and $r(l_b) \le \frac{\alpha}{2L'}$. This is possible by taking l_b small enough. Using that eigenvalues are 1-Lipschitz in the Hessian, we see for any $\mathbf{w}$ and arbitrary $\mathbf{w}^\star \in \mathcal{W}^\star$ that

  $$\left|\lambda_{\min}(\nabla^2 F(\mathbf{w}))\right| = \left|\lambda_{\min}(\nabla^2 F(\mathbf{w})) - \lambda_{\min}(\nabla^2 F(\mathbf{w}^\star))\right| \le \left\|\nabla^2 F(\mathbf{w}) - \nabla^2 F(\mathbf{w}^\star)\right\|_{\text{op}} \le L'\|\mathbf{w} - \mathbf{w}^\star\|.$$

  It follows for all $\mathbf{w}$ with $\|\mathbf{w} - \mathbf{w}^\star\| \le \frac{\alpha}{2L'}$, F is $\alpha/2$-strongly convex.

We next formally instantiate the corollaries of Theorem 12, Theorem 15 for the examples from Section 4. Directly applying Theorem 12, Theorem 15 for F satisfying the conditions of Example 3, Example 4, Example 5 implies the following.

Corollary 23 (Implications for Isoperimetry and Sampling) *Suppose F satisfies the conditions of Example 3, Example 4, or Example 5, and F also satisfies Assumption 3.1, Assumption 3.2. Then the following holds true:*

- *μ_β satisfies a PI with $\mathsf{C}_{\text{PI}} = O(\mathsf{C}_{\text{PI, LOCAL}})$ for $\beta = \Omega(d)$.*
- *Under the conditions of point 2 of Theorem 12, we also obtain a LSI for μ_β with $\mathsf{C}_{\text{LSI}} = O(S\beta\mathsf{C}_{\text{PI, LOCAL}})$ for $\beta = \Omega(d)$, where S is the second moment of μ_β.*
- *Suppose that F satisfies the conditions of Example 3, Example 4, or Example 5 outside some set $\mathcal{S}$. In this case, we obtain an $O((\mathsf{C}_{\text{PI, LOCAL}}), O(\mu_\beta(\mathcal{S}))$-WPI for μ_β for $\beta = \Omega(d)$.*

18. Which applies to a domain of $\mathbb{R}^d$ with convex boundary, see page 20, Bonnefont (2022).

- *As per Corollary 17, we can obtain a WPI for all these examples if $\mu_{\beta,\text{LOCAL}}$ does not satisfy Assumption 3.1 but instead satisfies a $(\mathsf{C}_{\text{WPI, LOCAL}}, \delta_{\text{LOCAL}})$-WPI.*
- *Via Corollary 19, Corollary 20, we obtain sampling guarantees polynomial in $\beta, d, \frac{1}{\varepsilon}$ for discrete-time LMC under Assumption 3.3, Assumption D.1.*

Our sampling results hold when F satisfies Assumption 3.1. While this or analogous conditions are necessary as per Remark 11, note convex F immediately satisfy Assumption 3.1. Thus taking $l_b = 1$ in Theorem 12, using the result of Payne and Weinberger (Payne and Weinberger, 1960) which states $\mathsf{C}_{\text{PI, LOCAL}} = O\big(\text{diam}(\mathcal{W}^\star)^2 + r(l_b)^2\big)$, and combining with Example 1, we directly obtain the following.

Corollary 24 (Sampling from non-smooth convex functions via LMC) *Suppose F is convex. Then*

$$\mathsf{C}_{\text{PI}}(\mu_\beta) = O\left(\text{diam}(\mathcal{W}^\star)^2 + r(l_b)^2 + \frac{1}{\beta}\right) \text{ for } \beta \geq \Omega\left(d + 4R^2 \vee \frac{2}{r_2^2}\right).$$

Furthermore, as a direct corollary of Corollary 20, we obtain results on sampling from particular log-concave measures (with the temperature restriction) where the potential is not smooth, similar to Lehec (2023). In fact, in some senses our results are stronger; those of Lehec (2023) (see Theorem 5) do not permit tail growth of F that is an arbitrary polynomial in $\|\mathbf{w}\|$.

A.4. Sampling Under a Stochastic Gradient Oracle

We can also use our results on a Log-Sobolev Inequality, in particular part 2 of Theorem 12 for F optimizable from all initializations, to show we can sample from μ_β when we only have a stochastic gradient oracle $\nabla f \approx \nabla F$. To the best of knowledge, the most recent guarantees in this setting are Das et al. (2023); Huang et al. (2024), where a variety of discretizations of (1) are considered. For the algorithms themselves, we refer the reader to these papers.

Under standard assumptions on bounded variance of a stochastic gradient oracle, to the best of our knowledge, the state-of-the-art guarantees for LMC in this setting are Theorems 4.1 and 4.2 of Huang et al. (2024). The results of Huang et al. (2024) state the following. Suppose F satisfies L-smoothness and μ_β satisfies a Log-Sobolev Inequality with constant C_{LSI}, and that f is written as a finite sum log-density. Then letting σ be an upper bound on the variance of the stochastic gradients ∇f, we can sample in TV-error ε from μ_β using $\tilde{O}\left(\frac{\beta^3 \mathsf{C}_{\text{LSI}}^3 d^{1/2} \min\{d + \beta^2\sigma^2, d^{1/2}\beta^2\sigma^2\}}{\varepsilon^2}\right)$ expected queries to the stochastic gradient oracle.

Combine this with the second part of our Theorem 12 for optimizable F. Recall $\beta = \Omega(d)$ for our results. Under the assumptions of the second part of Theorem 12, and that F is finite-sum and L-smooth, we obtain the following from Theorem 12:

- In the setting of Example 1: Here $\mathsf{C}_{\text{PI, LOCAL}} = O(1)$ and so $\mathsf{C}_{\text{LSI}}(\mu_\beta) = O(S\beta(1 + d/\beta) + d/\beta)$. We obtain a sampling guarantee in TV of $\tilde{O}\left(\frac{\beta^5 (S\beta(1+d/\beta)+d/\beta)^3 d^{1/2}\sigma^2}{\varepsilon^2}\right)$ for the algorithm given in Theorem 4.1 of Huang et al. (2024).
- In the setting of Example 2: Here $\mathsf{C}_{\text{PI, LOCAL}} = O(1/\beta)$ and so $\mathsf{C}_{\text{LSI}}(\mu_\beta) = O(S(1 + d/\beta) + d/\beta)$. We obtain a sampling guarantees in TV of $\tilde{O}\left(\frac{d\beta^2}{\varepsilon^2}\right)$ for the discretization (2) with ∇f used in place of ∇F, and $\tilde{O}\left(\frac{\beta^5 (S(1+d/\beta)+d/\beta)^3 d^{1/2}\sigma^2}{\varepsilon^2}\right)$ for the same algorithm from Theorem 4.1 of Huang et al. (2024).

Note if we also assume the standard dissipativity condition in Raginsky et al. (2017); Xu et al. (2018); Zou et al. (2021); Mou et al. (2022), by Lemma 1 of Raginsky et al. (2017), we can take $S = O(d/\beta)$ in the above.

Appendix B. Additional Background

B.1. Markov Semigroup Theory

We introduce the concept of the infinitesimal generator of a Markov process, which will make this exposition and our proofs much more natural. We give only what is needed for our work and refer the reader to Chewi (2024); Bakry et al. (2014) for more details.

Definition 25 *The infinitesimal generator of a Markov process* $\mathbf{w}(t)$ *is the operator* $\mathcal{L}$ *defined on all (sufficiently differentiable) functions* ϕ *by*

$$\mathcal{L}\phi(\mathbf{w}) = \lim_{t\to 0} \frac{\mathbb{E}[\phi(\mathbf{w}(t))] - \phi(\mathbf{w})}{t}.$$

This can be thought of as the instantaneous derivative of the Markov process in expectation. It is well-known that for the Langevin Diffusion (1), the generator takes the following form:

$$\mathcal{L}\phi(\mathbf{w}) = -\langle \beta\nabla F(\mathbf{w}), \nabla\phi(\mathbf{w})\rangle + \Delta\phi(\mathbf{w}). \tag{8}$$

For example, this calculation can be found in Example 1.2.4 of Chewi (2024).

We also need to introduce the idea of symmetry of the measure μ with respect to the stochastic process. In particular, we say μ is *symmetric* with respect to the Langevin Diffusion (1) if for all infinitely differentiable f, g,

$$\int f\mathcal{L}g\,\mathrm{d}\mu = \int \mathcal{L}f g\,\mathrm{d}\mu.$$

It is well-known that μ_β is symmetric. See e.g. Example 1.2.18 of Chewi (2024). This is used in Lemma 26.

B.2. The Proximal Sampler

Earlier we only discussed the discretization (2) of the Langevin Diffusion (1), which as shown in Chewi et al. (2024); Vempala and Wibisono (2019), succeeds in sampling from μ_β if μ_β satisfies an isoperimetric inequality. Another discretization of (1) that can sample from μ_β if μ_β satisfies an isoperimetric inequality is the *Proximal Sampler*, first introduced in Titsias and Papaspiliopoulos (2018); Lee et al. (2021). See Lee et al. (2021); Chen et al. (2022); Liang and Chen (2022a,b); Fan et al. (2023); Altschuler and Chewi (2024) for a variety of important developments on the proximal sampler. To the best of our knowledge, the state-of-the-art guarantees for the Proximal Sampler with exact gradients are in Altschuler and Chewi (2024), Fan et al. (2023). For state-of-the-art guarantees for the Proximal Sampler with stochastic gradients, see Huang et al. (2024). The Proximal Sampler is motivated by the Proximal Point Method in optimization, and works as follows: fix $h > 0$ and consider the joint distribution π on $\mathbb{R}^d \times \mathbb{R}^d$ defined as follows:

$$\pi(\mathbf{w}, \mathbf{w}') := \frac{1}{Z}\exp\left(-\beta F(\mathbf{w}) - \frac{1}{2h}\left\|\mathbf{w} - \mathbf{w}'\right\|^2\right).$$

Initialize $\mathbf{w}_0 \sim \pi_0$ and perform the following recursion between two sequences $\mathbf{w}_k$ (the samples of interest) and $\mathbf{w}'_k$ (an auxiliary sequence) for $k \geq 0$:

1. Sample $\mathbf{w}'_k \sim \pi^{\mathbf{w}'|\mathbf{w}}(\cdot|\mathbf{w}_k) = \mathcal{N}(\mathbf{w}_k, hI_d)$.
2. Sample the next iterate $\mathbf{w}_{k+1} \sim \pi^{\mathbf{w}|\mathbf{w}'}(\cdot|\mathbf{w}'_k)$.

Notice the second step is implementable if F is L-smooth for small enough $h \leq \frac{1}{2\beta L}$, as for such h, $\pi^{\mathbf{w}|\mathbf{w}'}(\cdot|\mathbf{w})$ is log-concave. In fact in Altschuler and Chewi (2024) and many other works on the proximal sampler, it is shown the Proximal Sampler is implementable with a *Proximal Oracle*, which given $\mathbf{w}' \in \mathbb{R}^d$, returns

$$\operatorname{argmin}_{\mathbf{w}\in\mathbb{R}^d}\left(F(\mathbf{w}) + \frac{1}{2h}\|\mathbf{w} - \mathbf{w}'\|^2\right).$$

A Proximal Oracle is implementable if F is smooth, as for small enough h, the above optimization problem is smooth and strongly convex. When we cite Theorems 5.3, 5.4 from Altschuler and Chewi (2024), we assume F is smooth.

B.3. The Tamed Unadjusted Langevin Algorithm

Here, we describe in detail the Weakly-Dissipative/Regularized Tamed Unadjusted Langevin Algorithm from Lytras and Mertikopoulos (2024). In recent years, works such as Lehec (2023); Lytras and Sabanis (2023); Lytras and Mertikopoulos (2024) have aimed to develop sampling algorithms that succeed beyond the relatively restrictive smoothness or Hölder continuity conditions in a variety of settings. As shown in 2.3 of Lytras and Mertikopoulos (2024), one needs to modify the sampling algorithm beyond (2) to sample from the Gibbs measure when F grows faster than a quadratic in $\|\mathbf{w}\|$. To our knowledge, the most general guarantees are in Lytras and Mertikopoulos (2024), and so we go with the results from there. The idea of these tamed sampling schemes is to split the gradient into two parts: one that grows at most linearly, and another part which we 'tame'. This allows for convergence results under far milder regularity conditions, Assumption 1 of Lytras and Mertikopoulos (2024), which we fully present in Assumption D.1, though we note Assumption D.1 is implied by Assumption A.1.

The Weakly-Dissipative Tamed Unadjusted Langevin Algorithm (wd-TULA) from their work gives an algorithm with more efficient guarantees under weak convexity of F or a LSI, and is defined as follows. Letting η denote the step size, we first let

$$f(\mathbf{w}) \coloneqq \beta\nabla F(\mathbf{w}) - \beta A_2\mathbf{w}(1 + \|\mathbf{w}\|^2)^{\frac{s_2}{2}-1}, f_\eta(\mathbf{w}) = \frac{f(\mathbf{w})}{1 + \sqrt{\eta}\|\mathbf{w}\|^{2s_3}},$$

where A_2, s_2, s_3 are defined in Assumption A.1. We then let

$$h_\eta(\mathbf{w}) \coloneqq \beta A_2\mathbf{w}(1 + \|\mathbf{w}\|^2)^{\frac{s_2}{2}-1} + f_\eta(\mathbf{w}),$$

and use $h_\eta(\mathbf{w})$ in place of $\beta\nabla F(\mathbf{w})$ in (2). That is, for standard d-dimensional normals ε_t,

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta h_\eta(\mathbf{w}_t) + \sqrt{2\eta}\varepsilon_t. \tag{9}$$

We use Theorem 2 of Lytras and Mertikopoulos (2024), which obtains a nonasymptotic polynomial-time guarantee for (9) under Assumption D.1 and a LSI for μ_β. The guarantee depends on the initialization $\mathsf{KL}(\pi_0\|\mu_\beta)$, but we argue in Lemma 46 that this can be controlled for appropriate π_0.

However, the Weakly-Dissipative Tamed Unadjusted Langevin Algorithm (wd-TULA) does not succeed when μ_β satisfies a PI. To this end, for large enough r (for example, $r = 4s_3 + 4$ is enough), we instead define (9) the same way as above, except F is replaced by a regularized version, $F(\mathbf{w}) + \frac{\lambda}{\beta}\|\mathbf{w}\|^{2r+2}$. That is, in defining $f(\mathbf{w})$ above, we take $\nabla\left(F(\mathbf{w}) + \frac{\lambda}{\beta}\|\mathbf{w}\|^{2r+2}\right)$ rather than $\nabla F(\mathbf{w})$. This yields the Regularized Tamed Unadjusted Langevin Algorithm (reg-TULA). In Theorem 3 of Lytras and Mertikopoulos (2024), reg-TULA was shown to succeed in sampling from μ_β under Assumption D.1 and a PI for μ_β. Again, we argue in Lemma 46 that the initialization error can be controlled for appropriate π_0.

Appendix C. Proof Ideas

Here, we sketch our proof; our full proofs are in Section D. We invite the reader interested in learning our proofs to first read this subsection, as we will build off the work here in Section D.

C.1. Proving a PI

The central idea is to prove a PI via the Lyapunov potential arising from optimization, a similar idea to Bakry et al. (2008). However, we modify their technique in a novel way to fully exploit local geometric properties implied by success of Gradient Descent, which gives us sharper quantitative control of the isoperimetric constant. Rather than building an ad-hoc Lyapunov potential from F, we instead utilize Φ as our potential in proving the functional inequality.

In our setting, recall we have a twice-differentiable and non-negative *Lyapunov function* $\Phi(\mathbf{w})$ such that

$$\langle \nabla\Phi(\mathbf{w}), \nabla F(\mathbf{w})\rangle \geq g(F(\mathbf{w}))$$

for a non-negative, monotonically increasing g with $g(x) \geq m'x - b'$, $g(0) = 0$, and $g(x) > 0$ for $x > 0$.

Recall the definition of the infinitesimal generator $\mathcal{L}$ of (1), defined as the following operator on any sufficiently differentiable test function ϕ:

$$\mathcal{L}\phi(\mathbf{w}) = \Delta\phi(w) - \langle \beta\nabla F(\mathbf{w}), \nabla\phi(\mathbf{w})\rangle.$$

Crucial to our analysis is the following Integration by Parts identity, which holds by reversibility of the Langevin Diffusion (1):

Lemma 26 (Theorem 1.2.14, Chewi (2024)) *For all functions f, g for which $\mathcal{L}f$, $\mathcal{L}g$ are defined,*

$$\int (-\mathcal{L})fg\mathrm{d}\mu_\beta = \int f(-\mathcal{L})g\mathrm{d}\mu_\beta = \int \langle \nabla f, \nabla g\rangle \mathrm{d}\mu_\beta.$$

For more background on the infinitesimal generator and the above identity, see Subsection B.1.

Now, we outline our argument. Following the discussion from Section 5, it remains to upper bound a term of the form $\int f(\mathbf{w})^2 \frac{g(F(\mathbf{w}))}{g(F(\mathbf{w}))+B}\mathrm{d}\mu_\beta$. We do so via Lemma 27, which crucially uses Lemma 26. Consider any $B > 0$, and let $h(\mathbf{w}) = g(F(\mathbf{w})) + B$. Lemma 27 gives

$$\int f(\mathbf{w})^2 \frac{g(F(\mathbf{w}))}{g(F(\mathbf{w})) + B}\mathrm{d}\mu_\beta \leq \frac{1}{\beta}\int \left(\|\nabla f(\mathbf{w})\|^2 + \frac{f(\mathbf{w})^2}{h(\mathbf{w})^2}\|\nabla\Phi(\mathbf{w})\|^2 - \frac{f(\mathbf{w})^2}{h(\mathbf{w})^2}\langle \nabla h(\mathbf{w}), \nabla\Phi(\mathbf{w})\rangle \right)\mathrm{d}\mu_\beta + \frac{1}{\beta}\int f(\mathbf{w})^2 \frac{|\Delta\Phi(\mathbf{w})|}{h(\mathbf{w})}\mathrm{d}\mu_\beta.$$

With Lemma 27 in hand and following the ideas from Section 5, we prove Theorem 12 by using Assumption 3.2 to upper bound

$$\frac{\|\nabla\Phi(\mathbf{w})\|^2}{h(\mathbf{w})^2} - \frac{\langle \nabla h(\mathbf{w}), \nabla\Phi(\mathbf{w})\rangle}{h(\mathbf{w})^2} \leq C',$$

for some $C' > 0$. Plugging this upper bound into the implication of Lemma 27 above and combining with the discussion from Section 5, rearranging and converting from f back to ψ eventually gives the desired PI. Using the 'tightening' technique of Cattiaux et al. (2010), we can strengthen the PI into a LSI for μ_β under the assumption of quadratic tail growth for F – which goes hand-in-hand with a LSI – and weak-convexity. Finally, once we have proved a PI or LSI, sampling from μ_β via LMC is known from the literature. This is fully detailed in Section D.

C.2. Proving a WPI

We also extend the Lyapunov technique to prove an WPI, which may be of independent interest. The idea is as follows: if (3) does not hold in $\mathcal{S}$ but otherwise holds in $\mathcal{S}^c$, instead consider arbitrary test function ψ and let $f = \psi - \alpha$ be defined exactly as in Section 5.

Considering any $B > 0$, for all $\mathbf{w} \in \mathcal{S}^c$, we have:

$$0 < g(F(\mathbf{w})) + B \leq \langle \nabla\Phi(\mathbf{w}), F(\mathbf{w})\rangle + B = -\frac{1}{\beta}\mathcal{L}\Phi(\mathbf{w}) + \frac{1}{\beta}\Delta\Phi(\mathbf{w}) + B \leq -\frac{1}{\beta}\mathcal{L}\Phi(\mathbf{w}) + \frac{1}{\beta}|\Delta\Phi(\mathbf{w})| + B.$$

Defining again $h(\mathbf{w}) = g(F(\mathbf{w})) + B > 0$, we obtain that

$$1 \leq \frac{1}{\beta}\cdot\frac{-\mathcal{L}\Phi}{h} + \frac{1}{\beta}\cdot\frac{|\Delta\Phi|}{h} + \frac{B}{h}.$$

Rather than integrating this inequality everywhere, we integrate it only where this holds, in $\mathcal{S}^c$. Multiplying the above by f^2 and integrating w.r.t. μ_β over $\mathcal{S}^c$ gives

$$\begin{aligned}
\int f^2 \mathrm{d}\mu_\beta &= \int_{\mathcal{S}} f^2 \mathrm{d}\mu_\beta + \int_{\mathcal{S}^c} f^2 \mathrm{d}\mu_\beta \\
&\leq \int_{\mathcal{S}} f^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\int_{\mathcal{S}^c} f^2 \frac{-\mathcal{L}\Phi}{h}\mathrm{d}\mu_\beta + \frac{1}{\beta}\int_{\mathcal{S}^c} f^2\frac{|\Delta\Phi|}{h}\mathrm{d}\mu_\beta + \int_{\mathcal{S}^c} f^2\frac{B}{h}\mathrm{d}\mu_\beta \\
&\leq \frac{1}{\beta}\int f^2\frac{-\mathcal{L}\Phi}{h}\mathrm{d}\mu_\beta + \frac{1}{\beta}\int f^2\frac{|\Delta\Phi|}{h}\mathrm{d}\mu_\beta + \int f^2\frac{B}{h}\mathrm{d}\mu_\beta + \left(\int_{\mathcal{S}} f^2\mathrm{d}\mu_\beta - \frac{1}{\beta}\int_{\mathcal{S}} f^2\frac{-\mathcal{L}\Phi}{h}\mathrm{d}\mu_\beta\right) \\
&\leq \frac{1}{\beta}\int f^2\frac{-\mathcal{L}\Phi}{h}\mathrm{d}\mu_\beta + \frac{1}{\beta}\int f^2\frac{|\Delta\Phi|}{h}\mathrm{d}\mu_\beta + \int f^2\frac{B}{h}\mathrm{d}\mu_\beta + \left(\int_{\mathcal{S}} f^2\mathrm{d}\mu_\beta + \frac{1}{\beta}\left|\int_{\mathcal{S}} f^2\frac{-\mathcal{L}\Phi}{h}\mathrm{d}\mu_\beta\right|\right).
\end{aligned}$$

The key difference is that the condition above not holding everywhere implies we picked up the 'error term'

$$\int_{\mathcal{S}} f^2\mathrm{d}\mu_\beta + \frac{1}{\beta}\left|\int_{\mathcal{S}} f^2\frac{-\mathcal{L}\Phi}{h}\mathrm{d}\mu_\beta\right|,$$

which we wish to relate to $\operatorname{osc}(\psi)$ to establish a WPI.

Notice for $\beta = \Omega(d)$, $\frac{1}{\beta}\left|\frac{-\mathcal{L}\Phi}{h}\right|$ can be controlled by a constant depending on problem-dependent parameters involving supremums over $\mathcal{S}$ (which is typically thought of as small).

Now we aim to see why f^2 can be related to $\operatorname{osc}(\psi)^2$. Indeed, since $f = \psi - \alpha$ where ψ is an expectation of ψ w.r.t a probability measure, namely $\mu_{\beta,\text{LOCAL}}$, we obtain that $|f| \leq \operatorname{osc}(\psi)$ pointwise. Consequently we can upper bound the error term by

$$\int_{\mathcal{S}} f^2\mathrm{d}\mu_\beta + \frac{1}{\beta}\left|\int_{\mathcal{S}} f^2\frac{-\mathcal{L}\Phi}{h}\mathrm{d}\mu_\beta\right| \leq \text{problem dependent parameters}\cdot\operatorname{osc}(\psi)^2\cdot\mu_\beta(\mathcal{S}).$$

Thus, rearranging the above, we obtain

$$\begin{aligned}
&\int f^2\cdot\frac{g(F(\mathbf{w}))}{h(\mathbf{w})}\mathrm{d}\mu_\beta \\
&\leq \frac{1}{\beta}\int f^2\frac{-\mathcal{L}\Phi}{h}\mathrm{d}\mu_\beta + \frac{1}{\beta}\int f^2\frac{|\Delta\Phi|}{h}\mathrm{d}\mu_\beta + \left(\int_{\mathcal{S}} f^2\mathrm{d}\mu_\beta + \frac{1}{\beta}\left|\int_{\mathcal{S}} f^2\frac{-\mathcal{L}\Phi}{h}\mathrm{d}\mu_\beta\right|\right) \\
&\leq \frac{1}{\beta}\int f^2\frac{-\mathcal{L}\Phi}{h}\mathrm{d}\mu_\beta + \frac{1}{\beta}\int f^2\frac{|\Delta\Phi|}{h}\mathrm{d}\mu_\beta + \text{problem dependent parameters}\cdot\operatorname{osc}(\psi)^2\cdot\mu_\beta(\mathcal{S}).
\end{aligned}$$

From here, we proceed similarly to our proof of the PI from earlier. Finally, to prove Corollary 17, rather than applying a PI for $\mu_{\beta,\text{LOCAL}}$ to upper bound $\int_{\mathcal{U}} f^2\mathrm{d}\mu_\beta$ from Assumption 3.1, apply the hypothesis that $\mu_{\beta,\text{LOCAL}}$ satisfies a WPI and use the same steps as above.

Appendix D. Proofs

In all of these proofs, we define $\mathcal{U} = \mathbb{B}(\mathcal{W}^\star, r(l_b))$ for l_b satisfying Assumption 3.1, as done in Section C.

D.1. Proof of Theorem 12

We first introduce the following Lemma.

Lemma 27 *Consider any Lyapunov function $\Phi(\cdot)$ and $g(\cdot)$ satisfying (3) for all $\mathbf{w} \in \mathbb{R}^d$. Then for any $B > 0$ and any test function f, we have*

$$\int f(\mathbf{w})^2 \frac{g(F(\mathbf{w}))}{g(F(\mathbf{w})) + B} \mathrm{d}\mu_\beta \le \frac{1}{\beta} \int \left(\|\nabla f(\mathbf{w})\|^2 + \frac{f(\mathbf{w})^2}{h(\mathbf{w})^2} \|\nabla \Phi(\mathbf{w})\|^2 - \frac{f(\mathbf{w})^2}{h(\mathbf{w})^2} \langle \nabla h(\mathbf{w}), \nabla \Phi(\mathbf{w}) \rangle \right) \mathrm{d}\mu_\beta + \frac{1}{\beta} \int f(\mathbf{w})^2 \frac{|\Delta \Phi(\mathbf{w})|}{h(\mathbf{w})} \mathrm{d}\mu_\beta. \tag{10}$$

Note when we apply Lemma 27, we will apply it with $\tilde{\Phi}$ in place of Φ and $\tilde{g}$ in place of g, where $\tilde{\Phi}, \tilde{g}$ are such that (3) holds with $\tilde{\Phi}$ in place of Φ and $\tilde{g}$ in place of g.

Proof of Lemma 27. By the condition (3), we obtain

$$g(F(\mathbf{w})) + B \le \langle \nabla \Phi(\mathbf{w}), F(\mathbf{w}) \rangle + B = -\frac{1}{\beta} \mathcal{L}\Phi(\mathbf{w}) + \frac{1}{\beta} \Delta \Phi(\mathbf{w}) + B \le -\frac{1}{\beta} \mathcal{L}\Phi(\mathbf{w}) + \frac{1}{\beta} |\Delta \Phi(\mathbf{w})| + B. \tag{11}$$

Denote $h(\mathbf{w}) := g(F(\mathbf{w})) + B > 0$. Therefore for any f, as $f^2 \ge 0$, we obtain

$$\begin{aligned}
\int f(\mathbf{w})^2 \mathrm{d}\mu_\beta &\le \int f(\mathbf{w})^2 \frac{-\frac{1}{\beta}\mathcal{L}\Phi(\mathbf{w}) + \frac{1}{\beta}|\Delta \Phi(\mathbf{w})| + B}{h(\mathbf{w})} \mathrm{d}\mu_\beta \\
&\le \frac{1}{\beta} \int f(\mathbf{w})^2 \frac{-\mathcal{L}\Phi(\mathbf{w})}{h(\mathbf{w})} \mathrm{d}\mu_\beta + \frac{1}{\beta} \int f(\mathbf{w})^2 \frac{|\Delta \Phi(\mathbf{w})|}{h(\mathbf{w})} \mathrm{d}\mu_\beta + \int f(\mathbf{w})^2 \frac{B}{h(\mathbf{w})} \mathrm{d}\mu_\beta.
\end{aligned}$$

For the first term, we use Lemma 26 in the second equality below to obtain

$$\begin{aligned}
\int f(\mathbf{w})^2 \frac{-\mathcal{L}\Phi(\mathbf{w})}{h(\mathbf{w})} \mathrm{d}\mu_\beta &= \int \frac{f(\mathbf{w})^2}{h(\mathbf{w})} \cdot -\mathcal{L}\Phi(\mathbf{w}) \mathrm{d}\mu_\beta \\
&= \int \left\langle \nabla \left(\frac{f(\mathbf{w})^2}{h(\mathbf{w})} \right), \nabla \Phi(\mathbf{w}) \right\rangle \mathrm{d}\mu_\beta \\
&= \int \left(\frac{2f(\mathbf{w})}{h(\mathbf{w})} \langle \nabla f(\mathbf{w}), \nabla \Phi(\mathbf{w}) \rangle - \frac{f(\mathbf{w})^2}{h(\mathbf{w})^2} \langle \nabla h(\mathbf{w}), \nabla \Phi(\mathbf{w}) \rangle \right) \mathrm{d}\mu_\beta \\
&\le \int \left(2 \left| \frac{f(\mathbf{w})}{h(\mathbf{w})} \right| \|\nabla f(\mathbf{w})\| \|\nabla \Phi(\mathbf{w})\| - \frac{f(\mathbf{w})^2}{h(\mathbf{w})^2} \langle \nabla h(\mathbf{w}), \nabla \Phi(\mathbf{w}) \rangle \right) \mathrm{d}\mu_\beta \\
&\le \int \|\nabla f(\mathbf{w})\|^2 + \frac{f(\mathbf{w})^2}{h(\mathbf{w})^2} \|\nabla \Phi(\mathbf{w})\|^2 - \frac{f(\mathbf{w})^2}{h(\mathbf{w})^2} \langle \nabla h(\mathbf{w}), \nabla \Phi(\mathbf{w}) \rangle \mathrm{d}\mu_\beta.
\end{aligned}$$

Combining the above two inequalities and rearranging gives

$$\int f(\mathbf{w})^2 \frac{g(F(\mathbf{w}))}{g(F(\mathbf{w})) + B} \mathrm{d}\mu_\beta \le \frac{1}{\beta} \int \left(\|\nabla f(\mathbf{w})\|^2 + \frac{f(\mathbf{w})^2}{h(\mathbf{w})^2} \|\nabla \Phi(\mathbf{w})\|^2 - \frac{f(\mathbf{w})^2}{h(\mathbf{w})^2} \langle \nabla h(\mathbf{w}), \nabla \Phi(\mathbf{w}) \rangle \right) \mathrm{d}\mu_\beta + \frac{1}{\beta} \int f(\mathbf{w})^2 \frac{|\Delta \Phi(\mathbf{w})|}{h(\mathbf{w})} \mathrm{d}\mu_\beta,$$

and this proves Lemma 27. ∎

Now, we prove Theorem 12.

Proof of Theorem 12. Our proof proceeds in three parts:

- Appropriately modifying Φ to make it more regular (which does *not* require additional regularity assumptions beyond those stated in Theorem 12).
- Using the Lyapunov function technique in a novel manner as sketched in Section C to prove a PI.
- Turning a PI into an LSI using established methods.

Part 1: Modifying Φ to introduce additional regularity. The first part of our proof is to show we can create a smooth (bounded Hessian eigenvalues) Lyapunov function $\tilde{\Phi}$ with that satisfies (3). The dependence on the allowed β and the resulting isoperimetric constants will in turn depend on $\tilde{\Phi}$. We emphasize this step is *only necessary when Φ is not smooth.*

First note without loss of generality we can take $m' \leftarrow \min\left(m', \frac{1}{2}\right)$. Also note we can without loss of generality replace g with a lower bound $\tilde{g}$ such that $\tilde{g}(0) = 0$, $\tilde{g}(x) > 0$ for $x > 0$, is increasing, and has exactly linear tail growth. We do so by constructing $\tilde{g}$ as follows. First define

$$x' = \frac{1}{m'}\left(g(r_2 R) + b'\right), \tag{12}$$

and notice that $m'x' - b' = g(r_2 R) > 0$. We now construct $\tilde{g}(\cdot)$ as follows:

- If $r_2 R \geq x'$, define:

$$\tilde{g}(x) = \begin{cases} \frac{1}{2}g(x) & \text{for } x \leq r_2 R \\ \text{smoothed version} & \text{for } x \in [r_2 R, r_2 R + \delta] \\ m'x - b' & \text{for } x \geq r_2 R + \delta \end{cases}$$

  for a small enough universal constant $\delta > 0$. By 'smoothed version' we just mean interpolating between the relevant two functions to preserve that $\tilde{g}(x)$ is differentiable and increasing while staying under the line $m'x - b'$, which we can easily see is possible because $m'x' - b' > \frac{1}{2}g(r_2 R) = \tilde{g}(r_2 R)$.

- Otherwise if $r_2 R < x'$, define:

$$\tilde{g}(x) = \begin{cases} \frac{1}{2}g(x) & \text{for } x \leq r_2 R \\ \text{smoothed version 1} & \text{for } x \in [r_2 R, r_2 R + \delta] \\ \frac{\frac{9}{10}(m'x'-b') - \frac{3}{4}g(r_2 R)}{x' - r_2 R}(x - r_2 R) + \frac{3}{4}g(r_2 R) & \text{for } x \in [r_2 R + \delta, x' - \delta] \\ \text{smoothed version 2} & \text{for } x \in [x' - \delta, x'] \\ m'x - b' & \text{for } x \geq x' \end{cases}$$

  for a small enough universal constant $\delta > 0$. Similarly as before, by 'smoothed version 1' we just mean interpolating between the relevant two functions to preserve that $\tilde{g}(x)$ is differentiable and increasing while staying under the line $\frac{\frac{9}{10}(m'x'-b') - \frac{3}{4}g(r_2 R)}{x' - r_2 R}(x - r_2 R) + \frac{3}{4}g(r_2 R)$, and likewise by 'smoothed version 2' we just mean interpolating between the relevant two functions to preserve that $\tilde{g}(x)$ is differentiable and increasing while staying under the line $m'x - b'$. This is possible because 1) $\frac{1}{2}g(r_2 R) < \frac{3}{4}g(r_2 R) < \frac{9}{10}g(r_2 R) < g(r_2 R) = m'x' - b'$, 2) $\frac{\frac{9}{10}(m'x'-b') - \frac{3}{4}g(r_2 R)}{x' - r_2 R}(x' - r_2 R) + \frac{3}{4}g(r_2 R) = \frac{9}{10}(m'x' - b') = \frac{9}{10}g(r_2 R) < g(r_2 R)$, and 3) $\frac{9}{10}(m'x' - b') - \frac{3}{4}g(r_2 R) = \frac{3}{20}g(r_2 R) > 0$. In particular, 1), 2) and 3) ensure we can always interpolate so that $\tilde{g}$ is increasing, and 2) also ensures that $\tilde{g}(x) \leq g(x)$.

Finally, take $\tilde{g}(x) \leftarrow r\tilde{g}(x)$ where

$$r = \min\left(1, \inf_{x \in [r_2 R, x']} \frac{x}{\tilde{g}(x)}\right). \tag{13}$$

Note $r > 0$ since $g(r_2 R) > 0$ and as $[r_2 R, x']$ is compact. These parameters also all behave in a dimension free way if m', b', r_2, R do.

In either case, the constructed $\tilde{g}(x)$ is increasing, differentiable, and has linear tail growth. In particular note $\tilde{g}(x) \geq r(m'(x - x') - b') = m'rx - r(m'x' + b')$. Moreover, by this construction, we can check that for $x \geq r_2 R$ we have $\tilde{g}(x) \leq x$, and for all $x \geq 0$ we have $g(x) \geq \tilde{g}(x)$. By Assumption 3.2, for all $\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R)^c$ we have $F(\mathbf{w}) \geq r_2 R$, therefore

$$\left\langle \frac{1}{r_1}(\mathbf{w} - \mathbf{w}^\star), \nabla F(\mathbf{w}) \right\rangle \geq F(\mathbf{w}) \geq \tilde{g}(F(\mathbf{w}))$$

outside $\mathbb{B}(\mathbf{w}^\star, R)$. Also, since for all x we have $g(x) \geq \tilde{g}(x)$, this implies for all $\mathbf{w}$ that

$$\langle \nabla \Phi(\mathbf{w}), \nabla F(\mathbf{w}) \rangle \geq g(F(\mathbf{w})) \geq \tilde{g}(F(\mathbf{w})).$$

Let $\Phi_2(\mathbf{w}) = \frac{1}{2r_1} \|\mathbf{w} - \mathbf{w}^\star\|^2 + M'$ where

$$M' := \sup_{\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R+1)} \Phi(\mathbf{w}). \tag{14}$$

Therefore, we have $\langle \nabla \Phi_2(\mathbf{w}), \nabla F(\mathbf{w}) \rangle \geq \tilde{g}(F(\mathbf{w}))$ outside $\mathbb{B}(\mathbf{w}^\star, R)$, and also that $\Phi_2(\mathbf{w}) \geq \Phi(\mathbf{w})$ on $\mathbb{B}(\mathbf{w}^\star, R+1)$. Furthermore, note the above construction of $\tilde{g}(x)$ is unnecessary if $g(x) = \lambda x$, by taking $\lambda = \min(\lambda, 1)$, which is the case in many of our examples e.g. Example 3, Example 5.

From here on out, if $g(x) = \lambda x$ for $\lambda \leq 1$ we define

$$m'_{\text{NEW}} = m', b'_{\text{NEW}} = b'. \tag{15}$$

Otherwise if the above construction of $\tilde{g}$ was needed we define

$$m'_{\text{NEW}} = m'r, b'_{\text{NEW}} = r(m'x' + b'), \tag{16}$$

where r, x' are defined as per (13), (12). Consequently we always have

$$\tilde{g}(x) \geq m'_{\text{NEW}} x - b'_{\text{NEW}}. \tag{17}$$

Now, we let $\chi(\mathbf{w}) \in [0, 1]$ be a bump function interpolating between $\mathbb{B}(\mathbf{w}^\star, R)$ and $\mathbb{B}(\mathbf{w}^\star, R+1)$ in the natural way, such that $\chi \equiv 0$ on $\mathbb{B}(\mathbf{w}^\star, R)$ and $\chi \equiv 1$ on $\mathbb{B}(\mathbf{w}^\star, R+1)^c$. In Lemma 47, we explicitly construct a $\chi(\mathbf{w})$ such that:

- $\chi(\mathbf{w})$ is differentiable to all orders.
- $\|\nabla \chi(\mathbf{w})\|, \|\nabla^2 \chi(\mathbf{w})\|_{\text{op}} \leq B$ where $B > 0$ is a universal constant.
- $\langle \nabla \chi(\mathbf{w}), \nabla F(\mathbf{w}) \rangle \geq 0$ for $\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R)^c \cap \mathbb{B}(\mathbf{w}^\star, R+1)$.

Now, define

$$\tilde{\Phi}(\mathbf{w}) := \chi(\mathbf{w})\Phi_2(\mathbf{w}) + (1 - \chi(\mathbf{w}))\Phi(\mathbf{w}).$$

We break into cases and show that $\tilde{\Phi}$ is still a valid Lyapunov function.

- For $\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R)$, as $\chi \equiv 0$ holds identically in this set, we have

$$\left\langle \nabla \tilde{\Phi}(\mathbf{w}), \nabla F(\mathbf{w}) \right\rangle \equiv \langle \nabla \Phi(\mathbf{w}), \nabla F(\mathbf{w}) \rangle \geq \tilde{g}(F(\mathbf{w})).$$

- For $\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R+1)^c$, as $\chi \equiv 1$ identically in this set, we have

$$\left\langle \nabla \tilde{\Phi}(\mathbf{w}), \nabla F(\mathbf{w}) \right\rangle = \langle \nabla \Phi_2(\mathbf{w}), \nabla F(\mathbf{w}) \rangle \geq \tilde{g}(F(\mathbf{w})).$$

- For $\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R)^c \cap \mathbb{B}(\mathbf{w}^\star, R+1)$, we have

$$\nabla \tilde{\Phi}(\mathbf{w}) = \chi(\mathbf{w}) \nabla \Phi_2(\mathbf{w}) + (1 - \chi(\mathbf{w})) \nabla \Phi(\mathbf{w}) + \nabla \chi(\mathbf{w}) \Phi_2(\mathbf{w}) - \nabla \chi(\mathbf{w}) \Phi(\mathbf{w}).$$

  This means

$$\begin{aligned} \left\langle \nabla \tilde{\Phi}(\mathbf{w}), \nabla F(\mathbf{w}) \right\rangle &= \chi(\mathbf{w}) \langle \nabla \Phi_2(\mathbf{w}), \nabla F(\mathbf{w}) \rangle + (1 - \chi(\mathbf{w})) \langle \nabla \Phi(\mathbf{w}), \nabla F(\mathbf{w}) \rangle \\ &\quad + (\Phi_2(\mathbf{w}) - \Phi(\mathbf{w})) \langle \nabla \chi(\mathbf{w}), \nabla F(\mathbf{w}) \rangle \\ &\geq (\chi(\mathbf{w}) + 1 - \chi(\mathbf{w})) \tilde{g}(F(\mathbf{w})) = \tilde{g}(F(\mathbf{w})). \end{aligned}$$

  The above uses that $\Phi_2(\mathbf{w}) \geq \Phi(\mathbf{w})$ for $\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R+1)$, and the property of χ that $\langle \nabla \chi(\mathbf{w}), \nabla F(\mathbf{w}) \rangle \geq 0$ for $\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R)^c \cap \mathbb{B}(\mathbf{w}^\star, R+1)$.

Therefore, for all $\mathbf{w} \in \mathbb{R}^d$ we have

$$\left\langle \nabla \tilde{\Phi}(\mathbf{w}), \nabla F(\mathbf{w}) \right\rangle \geq \tilde{g}(F(\mathbf{w})).$$

That is, $\tilde{\Phi}(\cdot)$, together with $\tilde{g}(\cdot)$, satisfies (3).

Moreover, we claim $\tilde{\Phi}$ is smooth. Note $\left\| \nabla^2 \Phi_2(\mathbf{w}) \right\|_{\text{op}} = \frac{1}{r_1}$ where r_1 was defined above. Let

$$L' = \sup_{\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R+1)} \rho_\Phi(\Phi(\mathbf{w})) \leq \rho_\Phi(M'), \tag{18}$$

where M' is as in (14).

- In $\mathbb{B}(\mathbf{w}^\star, R) \cup \mathbb{B}(\mathbf{w}^\star, R+1)^c$ we have $\left\| \nabla^2 \tilde{\Phi}(\mathbf{w}) \right\|_{\text{op}} \leq \max\left(L', \frac{1}{r_1}\right)$.
- In $\mathbb{B}(\mathbf{w}^\star, R)^c \cap \mathbb{B}(\mathbf{w}^\star, R+1)$, we can compute

$$\begin{aligned} \nabla^2 \tilde{\Phi}(\mathbf{w}) &= \nabla^2 \Phi(\mathbf{w}) + (\Phi_2(\mathbf{w}) - \Phi(\mathbf{w})) \nabla^2 \chi(\mathbf{w}) + \nabla^2 (\Phi_2(\mathbf{w}) - \Phi(\mathbf{w})) \chi(\mathbf{w}) \\ &\quad + 2 \nabla \chi(\mathbf{w}) \nabla (\Phi_2(\mathbf{w}) - \Phi(\mathbf{w}))^T. \end{aligned}$$

  By Triangle Inequality for operator norm and the inequality $\left\| \mathbf{a}\mathbf{b}^T \right\|_{\text{op}} \leq \|\mathbf{a}\| \|\mathbf{b}\|$, it follows that

$$\begin{aligned} &\left\| \nabla^2 \tilde{\Phi}(\mathbf{w}) \right\|_{\text{op}} \\ &\leq \left\| \nabla^2 \Phi(\mathbf{w}) \right\|_{\text{op}} + (|\Phi_2(\mathbf{w})| + |\Phi(\mathbf{w})|) \left\| \nabla^2 \chi(\mathbf{w}) \right\|_{\text{op}} + \left(\left\| \nabla^2 \Phi_2(\mathbf{w}) \right\|_{\text{op}} + \left\| \nabla^2 \Phi(\mathbf{w}) \right\|_{\text{op}} \right) \chi(\mathbf{w}) \\ &\qquad + 2 \| \nabla \chi(\mathbf{w}) \| \| \nabla (\Phi_2(\mathbf{w}) - \Phi(\mathbf{w})) \| \\ &\leq L' + B\left(\frac{(R+1)^2}{2r_1} + 2M' \right) + \left(\frac{1}{r_1} + L' \right) \cdot 1 + 2B\left(L' + \frac{R+1}{r_1} \right). \end{aligned}$$

Recalling L' from (18), define

$$\tilde{L} := \left\{L' + B\left(\frac{(R+1)^2}{2r_1} + 2M'\right) + \left(\frac{1}{r_1} + L'\right) + 2B\left(L' + \frac{R+1}{r_1}\right)\right\} \vee 2b'_{\text{NEW}} \vee 1, \tag{19}$$

where b'_{NEW} defines the linear univariate tail growth of $\tilde{g}$. Here, we recall the definitions of L' in (18), M' in (14), b'_{NEW} from (15) or (16) (whichever applies here), and B which is a universal constant coming from the construction of χ. Thus, $\tilde{\Phi}$ is $\tilde{L}$-smooth. Clearly $\tilde{\Phi}$ is non-negative as well.

Part 2: Proving a PI with the new Lyapunov function. Now we go back to our setup to prove a Poincaré Inequality. Consider any test function ψ. Let

$$f = \psi - \alpha \text{ where } \alpha = \frac{1}{\mu_\beta(\mathcal{U})}\int_{\mathcal{U}} \psi \mathrm{d}\mu_\beta. \tag{20}$$

Recall that $\tilde{\Phi}(\cdot)$ together with $\tilde{g}(\cdot)$ satisfies (3). Thus, applying Lemma 27 with $\tilde{\Phi}(\cdot)$, $\tilde{g}(\cdot)$, $B = \tilde{L} > 0$ and $h(\mathbf{w}) = \tilde{g}(F(\mathbf{w})) + \tilde{L}$ gives

$$\int f(\mathbf{w})^2 \frac{\tilde{g}(F(\mathbf{w}))}{\tilde{g}(F(\mathbf{w})) + \tilde{L}} \mathrm{d}\mu_\beta \le \frac{1}{\beta}\int \left(\|\nabla f(\mathbf{w})\|^2 + \frac{f(\mathbf{w})^2}{h(\mathbf{w})^2}\left\|\nabla\tilde{\Phi}(\mathbf{w})\right\|^2 - \frac{f(\mathbf{w})^2}{h(\mathbf{w})^2}\left\langle \nabla h(\mathbf{w}), \nabla\tilde{\Phi}(\mathbf{w})\right\rangle\right)\mathrm{d}\mu_\beta$$
$$+ \frac{1}{\beta}\int f(\mathbf{w})^2 \frac{\left|\Delta\tilde{\Phi}(\mathbf{w})\right|}{h(\mathbf{w})}\mathrm{d}\mu_\beta. \tag{21}$$

Step a: Upper bounding relevant terms using the construction of $\tilde{\Phi}$. We aim to upper bound the intermediate term in (21). Observe as $\tilde{g}$ is non-decreasing and non-negative,

$$\left\langle \nabla h(\mathbf{w}), \nabla\tilde{\Phi}(\mathbf{w})\right\rangle = \tilde{g}'(F(\mathbf{w}))\left\langle \nabla F(\mathbf{w}), \nabla\tilde{\Phi}(\mathbf{w})\right\rangle \ge \tilde{g}'(F(\mathbf{w}))\tilde{g}(F(\mathbf{w})) \ge 0.$$

Also observe by $\tilde{L}$-smoothness of $\tilde{\Phi}$ and using Lemma 42, because $\chi \in [0,1]$ and by definition of M',

$$\left\|\nabla\tilde{\Phi}(\mathbf{w})\right\|^2 \le 4\tilde{L}\tilde{\Phi}(\mathbf{w}) \le 4\tilde{L}\big(M' + \Phi_2(\mathbf{w})\big) = 4\tilde{L}\left(2M' + \frac{1}{2r_1}\|\mathbf{w} - \mathbf{w}^\star\|^2\right).$$

Therefore, as $g(x) \ge 0$, using the above implies

$$\frac{\left\|\nabla\tilde{\Phi}(\mathbf{w})\right\|^2 - \left\langle \nabla h(\mathbf{w}), \nabla\tilde{\Phi}(\mathbf{w})\right\rangle}{h(\mathbf{w})^2} \le \frac{\left\|\nabla\tilde{\Phi}(\mathbf{w})\right\|^2}{h(\mathbf{w})^2} \le \frac{4\tilde{L}\left(2M' + \frac{1}{2r_1}\|\mathbf{w} - \mathbf{w}^\star\|^2\right)}{h(\mathbf{w})^2}.$$

Furthermore recall that because $\tilde{g}(x) \ge \max(0, m'_{\text{NEW}}x - b'_{\text{NEW}})$, we have

$$h(\mathbf{w}) \ge \max\big(\tilde{L}, m'_{\text{NEW}}F(\mathbf{w}) - b'_{\text{NEW}} + \tilde{L}\big).$$

- If $\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R)$, using $\tilde{L}/2 \ge b'_{\text{NEW}}$, the above is clearly at most $\frac{8(R^2 + 4M'r_1)}{r_1\tilde{L}}$.
- Otherwise, using the second part of Assumption 3.2 and $\tilde{L}/2 \ge b'_{\text{NEW}}$, we have

$$\frac{4\tilde{L}\left(2M' + \frac{1}{2r_1}\|\mathbf{w} - \mathbf{w}^\star\|^2\right)}{\left(m'_{\text{NEW}}F(\mathbf{w}) - b'_{\text{NEW}} + \tilde{L}\right)^2} \le 4\tilde{L} \cdot \frac{\frac{1}{2r_1}\|\mathbf{w} - \mathbf{w}^\star\|^2 + 2M'}{r_2^2 m'^2_{\text{NEW}}\|\mathbf{w} - \mathbf{w}^\star\|^2 + \frac{\tilde{L}^2}{4}} \le \frac{2\tilde{L}}{r_1 r_2^2 m'^2_{\text{NEW}}} \vee \frac{32M'}{\tilde{L}}.$$

The last line uses the simple fact that $\frac{ta+b}{tc+d} \le \frac{a}{c} \vee \frac{b}{d}$ for all $t, a, b, c, d \ge 0$.

Define

$$C' := \frac{8\left(R^2 + 4M'r_1\right)}{r_1\tilde{L}} \vee \frac{2\tilde{L}}{r_1 r_2^2 m'^2_{\text{NEW}}} \vee \frac{32M'}{\tilde{L}}. \tag{22}$$

Here M' is from (14), $\tilde{L}$ is from (19), and m'_{NEW} is from (15) or (16) (whichever case applies here). Consequently the above proves that for any f, letting $h(\mathbf{w}) = \tilde{g}(F(\mathbf{w})) + \tilde{L}$, we have

$$\frac{\left\|\nabla\tilde{\Phi}(\mathbf{w})\right\|^2 - \left\langle\nabla h(\mathbf{w}), \nabla\tilde{\Phi}(\mathbf{w})\right\rangle}{h(\mathbf{w})^2} \le C'. \tag{23}$$

Step b: Using the Lyapunov method. Applying (23) in (21) and using $\tilde{L}$-smoothness of $\tilde{\Phi}$ and that $f^2 \ge 0$, we now have

$$\begin{aligned}
&\int f(\mathbf{w})^2 \frac{\tilde{g}(F(\mathbf{w}))}{\tilde{g}(F(\mathbf{w})) + \tilde{L}} \mathrm{d}\mu_\beta \\
&\le \frac{1}{\beta}\int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\int C' f(\mathbf{w})^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\int f(\mathbf{w})^2 \frac{\left|\Delta\tilde{\Phi}(\mathbf{w})\right|}{\tilde{g}(F(\mathbf{w})) + L}\mathrm{d}\mu_\beta \\
&\le \frac{1}{\beta}\int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\int C' f(\mathbf{w})^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\int f(\mathbf{w})^2 \frac{d\tilde{L}}{\tilde{g}(F(\mathbf{w})) + \tilde{L}}\mathrm{d}\mu_\beta \\
&\le \frac{1}{\beta}\int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\int f(\mathbf{w})^2 \left(d + C'\right)\mathrm{d}\mu_\beta.
\end{aligned}$$

Notice $\frac{\tilde{g}(t)}{\tilde{g}(t)+\tilde{L}}$ is non-decreasing as $\tilde{g}$ is non-decreasing. We thus obtain:

$$\begin{aligned}
&\int f(\mathbf{w})^2 \frac{\tilde{g}(l_b)}{\tilde{g}(l_b) + \tilde{L}} \mathrm{d}\mu_\beta \\
&= \int_{\mathcal{U}^c} f(\mathbf{w})^2 \frac{\tilde{g}(l_b)}{\tilde{g}(l_b) + \tilde{L}} \mathrm{d}\mu_\beta + \int_{\mathcal{U}} f(\mathbf{w})^2 \frac{\tilde{g}(l_b)}{\tilde{g}(l_b) + \tilde{L}} \mathrm{d}\mu_\beta \\
&\le \int f(\mathbf{w})^2 \frac{\tilde{g}(F(\mathbf{w}))}{\tilde{g}(F(\mathbf{w})) + \tilde{L}} \mathrm{d}\mu_\beta + \int_{\mathcal{U}} f(\mathbf{w})^2 \frac{\tilde{g}(l_b)}{\tilde{g}(l_b) + \tilde{L}} \mathrm{d}\mu_\beta \\
&\le \frac{1}{\beta}\int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\int f(\mathbf{w})^2 \left(d + C'\right)\mathrm{d}\mu_\beta + \frac{\tilde{g}(l_b)}{\tilde{g}(l_b) + \tilde{L}} \int_{\mathcal{U}} f(\mathbf{w})^2 \mathrm{d}\mu_\beta.
\end{aligned} \tag{24}$$

We now upper bound $\int_{\mathcal{U}} f(\mathbf{w})^2 \mathrm{d}\mu_\beta$. As $\mu_{\beta,\text{LOCAL}} := \mu_{\beta,\text{LOCAL}}(l_b)$ satisfies a Poincaré Inequality by Assumption 3.1, we have

$$\mathbb{V}_{\mu_{\beta,\text{LOCAL}}}(f) \le \mathsf{C}_{\text{PI, LOCAL}} \int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_{\beta,\text{LOCAL}}.$$

Using definition of variance and $\mu_{\beta,\text{LOCAL}}$ in the above, we obtain that

$$\frac{1}{\mu_\beta(\mathcal{U})}\int_{\mathcal{U}} f(\mathbf{w})^2 \mathrm{d}\mu_\beta - \frac{1}{\mu_\beta(\mathcal{U})^2}\left(\int_{\mathcal{U}} f(\mathbf{w})\mathrm{d}\mu_\beta\right)^2 \le \mathsf{C}_{\text{PI, LOCAL}} \cdot \frac{1}{\mu_\beta(\mathcal{U})}\int_{\mathcal{U}} \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta.$$

Recalling the definition of $f = \psi - \alpha$ for $\alpha = \frac{1}{\mu_\beta(\mathcal{U})}\int_{\mathcal{U}} \psi \mathrm{d}\mu_\beta$, we obtain from the above that

$$\int_{\mathcal{U}} f(\mathbf{w})^2 \mathrm{d}\mu_\beta \le \mathsf{C}_{\text{PI, LOCAL}} \int_{\mathcal{U}} \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta + \frac{1}{\mu_\beta(\mathcal{U})}\left(\int_{\mathcal{U}} f(\mathbf{w})\mathrm{d}\mu_\beta\right)^2$$

$$
\begin{aligned}
&\leq \mathsf{C}_{\text{PI, LOCAL}} \int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta + \frac{1}{\mu_\beta(\mathcal{U})}\left(\int_{\mathcal{U}}\left(\psi(\mathbf{w}) - \frac{1}{\mu_\beta(\mathcal{U})}\int_{\mathcal{U}} \psi(\mathbf{w})\mathrm{d}\mu_\beta\right)\mathrm{d}\mu_\beta\right)^2 \\
&= \mathsf{C}_{\text{PI, LOCAL}} \int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta + 0.
\end{aligned}
$$

Applying this in (24), we obtain

$$
\begin{aligned}
\int f(\mathbf{w})^2 \frac{\tilde{g}(l_b)}{\tilde{g}(l_b) + \tilde{L}} \mathrm{d}\mu_\beta \leq \frac{1}{\beta}\int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\int f(\mathbf{w})^2 \left(d + C'\right)\mathrm{d}\mu_\beta \\
+ \frac{\tilde{g}(l_b)}{\tilde{g}(l_b) + \tilde{L}} \cdot \mathsf{C}_{\text{PI, LOCAL}} \int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta.
\end{aligned}
$$

For $\beta \geq 2\left(1 + \frac{\tilde{L}}{\tilde{g}(l_b)}\right)(d + C') = \Omega(d)$, this gives

$$
\frac{\tilde{g}(l_b)}{2(\tilde{g}(l_b) + \tilde{L})}\int f(\mathbf{w})^2 \mathrm{d}\mu_\beta \leq \frac{1}{\beta}\int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta + \frac{\tilde{g}(l_b)}{\tilde{g}(l_b) + \tilde{L}} \cdot \mathsf{C}_{\text{PI, LOCAL}} \int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta.
$$

Rearranging this inequality and converting back to ψ, recalling the definition of variance, and noting $\nabla f = \nabla \psi$ gives:

$$
\begin{aligned}
\mathbb{V}_{\mu_\beta}[\psi] &\leq \int (\psi - \alpha)^2 \mathrm{d}\mu_\beta \\
&= \int f^2 \mathrm{d}\mu_\beta \\
&\leq \left(2\mathsf{C}_{\text{PI, LOCAL}} + \frac{2}{\beta}\left(1 + \frac{\tilde{L}}{\tilde{g}(l_b)}\right)\right)\int \|\nabla f\|^2 \mathrm{d}\mu_\beta \\
&= \left(2\mathsf{C}_{\text{PI, LOCAL}} + \frac{2}{\beta}\left(1 + \frac{\tilde{L}}{\tilde{g}(l_b)}\right)\right)\int \|\nabla \psi\|^2 \mathrm{d}\mu_\beta.
\end{aligned}
$$

Recalling ψ is an arbitrary test function, this shows that μ_β satisfies a Poincaré Inequality with a Poincaré constant of

$$
2\mathsf{C}_{\text{PI, LOCAL}} + \frac{2}{\beta}\left(1 + \frac{\tilde{L}}{\tilde{g}(l_b)}\right) \text{ for } \beta \geq 2\left(1 + \frac{\tilde{L}}{\tilde{g}(l_b)}\right)(d + C'), \tag{25}
$$

where $\tilde{L}$ is defined in (19) and C' is defined in (22).

Part 3: Proving a Log-Sobolev Inequality. With the above PI in hand, and under the relevant conditions given in Theorem 12, we use the following result of Cattiaux et al. (2010) in the form given by Proposition 15 from Raginsky et al. (2017) to prove an LSI.

Theorem 28 (Proposition 15, Raginsky et al. (2017)) *Suppose the following conditions hold:*

1. *There exists constants $\kappa, \gamma > 0$ and a twice continuously differentiable function $V : \mathbb{R}^d \to [1, \infty)$ such that for all $\mathbf{w} \in \mathbb{R}^d$,*

$$
\frac{\mathcal{L}V(\mathbf{w})}{V(\mathbf{w})} \leq \kappa - \gamma\|\mathbf{w}\|^2.
$$

2. *μ_β satisfies a Poincaré Inequality with constant C_{PI}.*

3. *There exists some constant $K \geq 0$ such that $\nabla^2 F \geq -K$.*

Then, for any $\delta > 0$. μ_β *satisfies a Log-Sobolev Inequality with* $\mathsf{C}_{\text{LSI}} = C_1 + (C_2 + 2)\mathsf{C}_{\text{PI}}$, *where*

$$C_1 := \frac{2}{\gamma}\left(\frac{1}{\delta} + \frac{\beta K}{2}\right) + \delta \quad \textit{and} \quad C_2 := \frac{2}{\gamma}\left(\frac{1}{\delta} + \frac{\beta K}{2}\right)\left(\kappa + \gamma \int_{\mathbb{R}^d} \|\mathbf{w}\|^2 \mathrm{d}\mu_\beta\right).$$

Use $V(\mathbf{w}) = e^{\tilde{\Phi}(\mathbf{w})}$ in Theorem 28. Condition 2 in Theorem 28 follows from the above part, and condition 3 in Theorem 28 is trivially satisfied with $K = L$ by our condition on weak convexity of F. For condition 1, let $V(\mathbf{w}) = e^{\tilde{\Phi}(\mathbf{w})} \geq 1$. Compute

$$\nabla V(\mathbf{w}) = e^{\tilde{\Phi}(\mathbf{w})}\nabla\tilde{\Phi}(\mathbf{w}), \Delta\tilde{\Phi}(\mathbf{w}) = e^{\tilde{\Phi}(\mathbf{w})}\left(\Delta\tilde{\Phi}(\mathbf{w}) + \left\|\nabla\tilde{\Phi}(\mathbf{w})\right\|^2\right).$$

Therefore,

$$\begin{aligned}\frac{\mathcal{L}V(\mathbf{w})}{V(\mathbf{w})} &= \frac{V(\mathbf{w})\left(\Delta\tilde{\Phi}(\mathbf{w}) + \left\|\nabla\tilde{\Phi}(\mathbf{w})\right\|^2 - \left\langle\beta\nabla F(\mathbf{w}), \nabla\tilde{\Phi}(\mathbf{w})\right\rangle\right)}{V(\mathbf{w})} \\ &= \Delta\tilde{\Phi}(\mathbf{w}) + \left\|\nabla\tilde{\Phi}(\mathbf{w})\right\|^2 - \left\langle\beta\nabla F(\mathbf{w}), \nabla\tilde{\Phi}(\mathbf{w})\right\rangle.\end{aligned}$$

We now upper bound the above. Recall we showed $\tilde{\Phi}(\mathbf{w})$ is $\tilde{L}$ smooth, hence $\Delta\tilde{\Phi}(\mathbf{w}) \leq d\tilde{L}$. Now we break into cases:

- Consider $\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R+1)$. Recall for such $\mathbf{w}$, $\left\|\nabla\tilde{\Phi}(\mathbf{w})\right\| \leq L'$. Also recall $\left\langle\nabla F(\mathbf{w}), \nabla\tilde{\Phi}(\mathbf{w})\right\rangle \geq \tilde{g}(F(\mathbf{w})) \geq 0$. Thus in this case
$$\frac{\mathcal{L}V(\mathbf{w})}{V(\mathbf{w})} \leq d\tilde{L} + L'.$$

- Consider $\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R+1)^c$. Now, $\left\|\nabla\tilde{\Phi}(\mathbf{w})\right\| = \frac{1}{r_1}\|\mathbf{w} - \mathbf{w}^\star\|$. Also recall $\left\langle\nabla F(\mathbf{w}), \nabla\tilde{\Phi}(\mathbf{w})\right\rangle \geq \tilde{g}(F(\mathbf{w}))$. By construction of $\tilde{g}$, we have $\tilde{g}(x) \geq m'_{\text{NEW}}x - b'_{\text{NEW}}$ (recall (17)). Hence, by conditions on the growth of F in this part, we obtain
$$\left\langle\nabla F(\mathbf{w}), \nabla\tilde{\Phi}(\mathbf{w})\right\rangle \geq \tilde{g}(F(\mathbf{w})) \geq m'_{\text{NEW}}(m\|\mathbf{w}\|^2 - b) - b'_{\text{NEW}} = mm'_{\text{NEW}}\|\mathbf{w}\|^2 - (bm'_{\text{NEW}} + b'_{\text{NEW}}).$$
Thus in this case, we have
$$\frac{\mathbf{L}V(\mathbf{w})}{V(\mathbf{w})} \leq d\tilde{L} + \frac{1}{r_1^2}\|\mathbf{w} - \mathbf{w}^\star\|^2 - \beta\left(mm'_{\text{NEW}}\|\mathbf{w}\|^2 - (bm'_{\text{NEW}} + b'_{\text{NEW}})\right).$$

Doing casework based on the above cases and with one application of Young's Inequality, we see that when $\beta \geq \frac{4}{r_1^2 m}$, condition 1 is of Theorem 28 is satisfied with

$$\begin{aligned}\kappa &= d\tilde{L} + L' + \frac{2}{r_1^2}\|\mathbf{w}^\star\|^2 + \beta(bm'_{\text{NEW}} + b'_{\text{NEW}}) + \frac{\beta mm'_{\text{NEW}}}{2}(R+1)^2, \\ \gamma &= \frac{\beta mm'_{\text{NEW}}}{2}.\end{aligned}$$

Choose $\delta = \frac{1}{\sqrt{\gamma}}$. As $\beta \geq 2$, we can check

$$C_1 = \frac{4}{mm'_{\text{NEW}}\beta}\left(\sqrt{\frac{\beta mm'_{\text{NEW}}}{2}} + \frac{\beta L}{2}\right) + \sqrt{\frac{2}{\beta mm'_{\text{NEW}}}} \leq \frac{4L+3}{2mm'_{\text{NEW}}} + \frac{3}{2},$$

$$
\begin{aligned}
C_2 &= 2\left(\sqrt{\gamma} + \frac{\beta L}{2}\right)\left(\frac{\kappa}{\gamma} + S\right) \\
&\le 2\left(\sqrt{\frac{\beta m m'_{\text{NEW}}}{2}} + \frac{\beta L}{2}\right) \\
&\qquad \cdot \left((R+1)^2 + \frac{2(b'_{\text{NEW}} + b m'_{\text{NEW}})}{m m'_{\text{NEW}}} + \frac{4}{\beta m m'_{\text{NEW}} r_1^2}\|\mathbf{w}^\star\|^2 + \frac{2\left(d\tilde{L} + L'\right)}{\beta m m'_{\text{NEW}}} + S\right).
\end{aligned}
$$

Using $\beta \ge 2$, and our earlier upper bound on C_{PI}, this yields a Log-Sobolev constant of

$$
\begin{aligned}
\mathsf{C}_{\text{LSI}} &\le C_1 + (C_2 + 2)\mathsf{C}_{\text{PI}} \\
&\le \frac{4L+3}{2 m m'_{\text{NEW}}} + \frac{3}{2} \\
&\quad + 4\left(1 + \left\{L + \sqrt{m m'_{\text{NEW}}}\right\} \cdot \left\{(R+1)^2 + 2\left(\frac{b'_{\text{NEW}}}{m m'_{\text{NEW}}} + \frac{b}{m}\right) + \frac{4}{\beta m m'_{\text{NEW}} r_1^2}\|\mathbf{w}^\star\|^2 + \frac{2\left(d\tilde{L} + L'\right)}{\beta m m'_{\text{NEW}}} + S\right\}\right) \\
&\qquad \cdot \left(\left\{1 + \frac{\tilde{L}}{\tilde{g}(l_b)}\right\} + \beta \mathsf{C}_{\text{PI, LOCAL}}\right),
\end{aligned} \tag{26}
$$

for $\beta \ge 2\left(1 + \frac{\tilde{L}}{\tilde{g}(l_b)}\right)(d + C') \vee \frac{4}{r_1^2 m} \ge 2$. Again, in the above, $\tilde{L}$ comes from (19), C' comes from (22), L' comes from (18), and $m'_{\text{NEW}}, b'_{\text{NEW}}$ are as per (15) or (16), whichever case is appropriate. This proves the desired LSI. ■

Remark 29 *Notice in the above proof, we did not use Assumption 1.1 on F, hence the statement of Theorem 12. Note also that this proof establishes a PI from optimizability almost everywhere (w.r.t. Lebesgue measure ν), since μ is absolutely continuous with respect to ν.*

Remark 30 *We note when Φ is L-smooth to begin with (for example, $L = 2$ when $\Phi(\mathbf{w}) = \|\mathbf{w} - \mathbf{w}^\star\|^2$, which holds in the Linearizable example Example 5), the construction of $\tilde{g}$ and $\tilde{\Phi}$ is unnecessary. We can just use Φ instead of $\tilde{\Phi}$, and in the above guarantees from (25), (26), using Lemma 42 we have*

$$
\tilde{L} = L \vee 2b', M' = 0, C' = \frac{8R^2}{\min(1/2, r_1)\tilde{L}} \vee \frac{2\tilde{L}}{\min(1/2, r_1) r_2^2 m'^2_{\text{NEW}}}. \tag{27}
$$

For example, in this case we obtain μ_β satisfies a Poincaré Inequality with a Poincaré constant of

$$
\mathsf{C}_{\text{PI}} = 2\mathsf{C}_{\text{PI, LOCAL}} + \frac{2}{\beta}\left(1 + \frac{L \vee 2b'}{g(l_b)}\right) \text{ for } \beta \ge 2\left(1 + \frac{L \vee 2b'}{g(l_b)}\right)\left(d + \frac{8R^2}{\min(1/2, r_1)\tilde{L}} \vee \frac{2\tilde{L}}{\min(1/2, r_1) r_2^2 m'^2_{\text{NEW}}}\right).
$$

We similarly obtain a cleaner and tighter bound for C_{LSI} plugging the expressions from (27) back into (26). Also note the construction of $\tilde{g}$ is unnecessary if $g(x) = \lambda x$ for $\lambda \le 1$; if this is the case, we can just take $m'_{\text{NEW}} = \lambda$, $b'_{\text{NEW}} = 0$.

Remark 31 *By tracking the proof, we see that if Assumption 3.2 holds, it suffices to have Φ, F, g satisfy (3) inside $\mathbb{B}(\mathbf{w}^\star, R+1)$. This is because in our construction of $\tilde{g}$, we did not change R. After this, in our construction of $\tilde{\Phi}$, we only need the condition from Assumption 3.2 outside $\mathbb{B}(\mathbf{w}^\star, R+1)$. After our construction of $\tilde{\Phi}$, the condition (3) is no longer used in the proof.*

Remark 32 *Consider a canonical example of non-convex, optimizable F: when F is λ-Linearizable (Kale et al., 2021; Kleinberg et al., 2018; De Sa et al., 2022; Hinder et al., 2020). For simplicity say $\lambda \le 1$. Thus*

Definition 2 holds with $\Phi = \|\mathbf{w} - \mathbf{w}^\star\|^2$ (which is 2-smooth) and $g(x) = \lambda x$. For

$$\beta \geq 2\left(1 + \frac{2}{\lambda l_b}\right)\left(d + \frac{8R^2}{\min(r_1, 1/2)} \vee \frac{4}{\lambda^2 \min(r_1, 1/2) r_2^2}\right), \tag{28}$$

Theorem 12 gives a PI. Note as Assumption 1.1 is not needed for F, no regularity assumptions are placed on F. Also note the construction of $\tilde{g}$ is unnecessary here, hence we can just take $m'_{\text{NEW}} = \lambda$, $b'_{\text{NEW}} = 0$.

The concurrent work Gong et al. (2024); Chewi and Stromme (2024) only consider PŁ functions, which is not a natural parametrization for this example. Both approaches also do not yield a PI without further assumptions on F. Examining Lemma 3.3 of Gong et al. (2024), they require $\beta \geq \frac{4dL}{g_0^2}$ where g_0 is a lower bound on the gradients outside $\mathcal{W}^\star$ and L is defined in their Assumption 4 and is analogous to the Lipschitz constant of the Hessian near $\mathcal{W}^\star$. Chewi and Stromme (2024) requires an upper bound on the Laplacian, which often scales with d, even in the standard setting when F is L-smooth and so $\Delta F \leq dL$. Following their approach to derive a PI, one needs $\beta\|\nabla F\|^2 \geq dL$ outside $\mathbf{w}^\star$ (see their page 10).

In this Linearizable setting under Assumption 3.2, all we can obtain for generic F is $\|\nabla F(\mathbf{w})\| \geq r_1 r_2 \wedge \frac{r_1 \lambda l_b}{R}$ outside $\mathcal{W}^\star$. Thus the techniques of Gong et al. (2024); Chewi and Stromme (2024) require

$$\beta \geq d(L \wedge L')\left(\frac{1}{r_1^2 r_2^2} \vee \frac{R^2}{\lambda^2 l_b^2 r_1^2}\right).$$

Often r_1, r_2 could be quite small and R is quite large; these costly terms are multiplied by the dimension d in the requirement for inverse temperature. This is not the case using our result Theorem 12 to obtain our inverse temperature requirement (28).

D.2. Proof of Weak Poincaré Inequality Results Theorem 15, Corollary 17

Proof of Theorem 15. For the rest of the proof, borrow the same notation as in the proof in Subsection D.1. Consider any test function ψ. As in (20), let

$$f = \psi - \alpha \text{ where } \alpha = \frac{1}{\mu_\beta(\mathcal{U})} \int_{\mathcal{U}} \psi \mathrm{d}\mu_\beta.$$

First, recall we can preserve Definition 2 by replacing Φ with $\tilde{\Phi}$ and g with $\tilde{g}$, as done in Part 1 of the proof in Subsection D.1. By the work there, which was all done *pointwise*, the resulting $\tilde{\Phi}$ still satisfies Definition 2, but *now only for all* $\mathbf{w} \in \mathcal{S}^c$. That is, we have

$$\left\langle \nabla\tilde{\Phi}(\mathbf{w}), \nabla F(\mathbf{w})\right\rangle \geq \tilde{g}(F(\mathbf{w})) \text{ for all } \mathbf{w} \in \mathcal{S}^c. \tag{29}$$

Moreover, the construction of $\tilde{\Phi}$ there using Assumption 3.2 ensures $\tilde{\Phi}$ satisfies $\left\langle \nabla\tilde{\Phi}(\mathbf{w}), \nabla F(\mathbf{w})\right\rangle \geq \tilde{g}(F(\mathbf{w}))$ for all $\mathbf{w} \in \mathbb{B}(\mathbf{w}, R+1)^c$. Thus, $\tilde{\Phi}$ does not satisfy Definition 2 only for $\mathbf{w} \in \mathcal{S} \cap \mathbb{B}(\mathbf{w}^\star, R+1)$, so we assume from now on that $\mathcal{S} \subseteq \mathbb{B}(\mathbf{w}^\star, R+1)$ (equivalently, take $\mathcal{S} \leftarrow \mathcal{S} \cap \mathbb{B}(\mathbf{w}^\star, R+1)$). The verification of the smoothness of $\tilde{\Phi}$ did not use optimizability, and so we know that $\tilde{\Phi}$ is $\tilde{L}$-smooth over all of $\mathbb{R}^d$, where $\tilde{L}$ is again defined as in (19).

Next, let

$$B = \tilde{L} \vee G_F G_\Phi \geq 1, \tag{30}$$

where again $\tilde{L}$ is from (19), and where we define

$$G_F := \sup_{\mathbf{w} \in \mathcal{S}} \|\nabla F(\mathbf{w})\| \leq L_F R, G_\Phi := \sup_{\mathbf{w} \in \mathcal{S}} \left\|\nabla\tilde{\Phi}(\mathbf{w})\right\| \leq \rho_\Phi(M'), \tag{31}$$

Here we define $M_F = \sup_{\mathbf{w}\in\mathbb{B}(\mathbf{w}^\star,R+1)} F(\mathbf{w})$ and upper bound

$$\left\|\nabla^2 F(\mathbf{w})\right\|_{\mathrm{op}} \le \rho_F(M_F) := L_F,$$

where we use Assumption 1.1 for F and that $\mathcal{S} \subseteq \mathbb{B}(\mathbf{w}^\star, R+1)$. Thus, the following holds for all $\mathbf{w} \in \mathcal{S}^c$:

$$0 < \tilde{g}(F(\mathbf{w})) + B \le \langle \nabla\tilde{\Phi}(\mathbf{w}), F(\mathbf{w})\rangle + B = -\frac{1}{\beta}\mathcal{L}\tilde{\Phi}(\mathbf{w}) + \frac{1}{\beta}\Delta\tilde{\Phi}(\mathbf{w}) + B \le -\frac{1}{\beta}\mathcal{L}\tilde{\Phi}(\mathbf{w}) + \frac{1}{\beta}\left|\Delta\tilde{\Phi}(\mathbf{w})\right| + B.$$

Define $h(\mathbf{w}) = \tilde{g}(F(\mathbf{w})) + B$. Thus for all $\mathbf{w} \in \mathcal{S}^c$,

$$1 \le -\frac{1}{\beta}\cdot\frac{\mathcal{L}\tilde{\Phi}(\mathbf{w})}{h} + \frac{1}{\beta}\cdot\frac{\left|\Delta\tilde{\Phi}(\mathbf{w})\right|}{h} + \frac{B}{h}.$$

Thus, as $f^2 \ge 0$, we obtain

$$\begin{aligned}
\int f^2 \mathrm{d}\mu_\beta &= \int_{\mathcal{S}} f^2 \mathrm{d}\mu_\beta + \int_{\mathcal{S}^c} f^2 \mathrm{d}\mu_\beta \\
&\le \int_{\mathcal{S}} f^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\int_{\mathcal{S}^c} f^2 \frac{-\mathcal{L}\tilde{\Phi}}{h}\mathrm{d}\mu_\beta + \frac{1}{\beta}\int_{\mathcal{S}^c} f^2\frac{\left|\Delta\tilde{\Phi}\right|}{h}\mathrm{d}\mu_\beta + \int_{\mathcal{S}^c} f^2\frac{B}{h}\mathrm{d}\mu_\beta \\
&\le \frac{1}{\beta}\int f^2\frac{-\mathcal{L}\tilde{\Phi}}{h}\mathrm{d}\mu_\beta + \frac{1}{\beta}\int f^2\frac{\left|\Delta\tilde{\Phi}\right|}{h}\mathrm{d}\mu_\beta + \int f^2\frac{B}{h}\mathrm{d}\mu_\beta + \left(\int_{\mathcal{S}} f^2\mathrm{d}\mu_\beta - \frac{1}{\beta}\int_{\mathcal{S}} f^2\frac{-\mathcal{L}\tilde{\Phi}}{h}\mathrm{d}\mu_\beta\right) \\
&\le \frac{1}{\beta}\int f^2\frac{-\mathcal{L}\tilde{\Phi}}{h}\mathrm{d}\mu_\beta + \frac{1}{\beta}\int f^2\frac{\left|\Delta\tilde{\Phi}\right|}{h}\mathrm{d}\mu_\beta + \int f^2\frac{B}{h}\mathrm{d}\mu_\beta + \left(\int_{\mathcal{S}} f^2\mathrm{d}\mu_\beta + \frac{1}{\beta}\left|\int_{\mathcal{S}} f^2\frac{-\mathcal{L}\tilde{\Phi}}{h}\mathrm{d}\mu_\beta\right|\right).
\end{aligned}$$

The last term in parantheses is now our error term. The first three terms will be controlled analogously to Subsection D.1. Namely, the same application of Integration by Parts as in Lemma 27, which never uses the optimizability condition, yields

$$\int f^2\frac{-\mathcal{L}\tilde{\Phi}}{h}\mathrm{d}\mu_\beta \le \int\left(\|\nabla f\|^2 + \frac{f^2}{h^2}\left\|\nabla\tilde{\Phi}\right\|^2 - \frac{f^2}{h^2}\langle\nabla h, \nabla\tilde{\Phi}\rangle\right)\mathrm{d}\mu_\beta.$$

Substituting this inequality in the above, we obtain in the same way as with (10) that

$$\begin{aligned}
\int f(\mathbf{w})^2\frac{\tilde{g}(F(\mathbf{w}))}{\tilde{g}(F(\mathbf{w}))+B}\mathrm{d}\mu_\beta \le \frac{1}{\beta}\int&\left(\|\nabla f(\mathbf{w})\|^2 + \frac{f(\mathbf{w})^2}{h(\mathbf{w})^2}\left\|\nabla\tilde{\Phi}(\mathbf{w})\right\|^2 - \frac{f(\mathbf{w})^2}{h(\mathbf{w})^2}\langle\nabla h(\mathbf{w}), \nabla\tilde{\Phi}(\mathbf{w})\rangle\right)\mathrm{d}\mu_\beta \\
&+ \frac{1}{\beta}\int f(\mathbf{w})^2\frac{\left|\Delta\tilde{\Phi}(\mathbf{w})\right|}{h(\mathbf{w})}\mathrm{d}\mu_\beta \\
&+ \left(\int_{\mathcal{S}} f(\mathbf{w})^2\mathrm{d}\mu_\beta + \frac{1}{\beta}\left|\int_{\mathcal{S}} f(\mathbf{w})^2\frac{-\mathcal{L}\tilde{\Phi}(\mathbf{w})}{h(\mathbf{w})}\mathrm{d}\mu_\beta\right|\right).
\end{aligned} \tag{32}$$

As discussed in Section C, we picked up the 'error term' $\int_{\mathcal{S}} f(\mathbf{w})^2\mathrm{d}\mu_\beta + \frac{1}{\beta}\left|\int_{\mathcal{S}} f(\mathbf{w})^2\frac{-\mathcal{L}\tilde{\Phi}(\mathbf{w})}{h(\mathbf{w})}\mathrm{d}\mu_\beta\right|$.

Step a. Now, we follow Part 2, Step a, Subsection D.1 to upper bound the first term in the right hand side above. Note for $\mathbf{w} \in \mathcal{S}^c$, we still have (23) for such $\mathbf{w}$, as the proof of (23) only used optimizability pointwise. Otherwise, consider $\mathbf{w} \in \mathcal{S}$. Let

$$G' = \sup_{t\in\mathbb{R}}|g'(t)|. \tag{33}$$

Note this is dimension free and has no F-dependence. Note by choice of $h(\mathbf{w})$,

$$-\langle \nabla h(\mathbf{w}), \nabla\tilde{\Phi}(\mathbf{w})\rangle \le \tilde{g}'(F(\mathbf{w}))\|\nabla F(\mathbf{w})\|\|\nabla\tilde{\Phi}(\mathbf{w})\| \le G'\Big(\|\nabla F(\mathbf{w})\|^2 + \|\nabla\tilde{\Phi}(\mathbf{w})\|^2\Big).$$

Furthermore recall that in Part 2, Step a of Subsection D.1, without using optimizability of F, it was established that $\frac{\|\nabla\tilde{\Phi}\|^2}{h(\mathbf{w})^2} \le C'$, where C' was defined in (23). Thus,

$$\frac{\|\nabla\tilde{\Phi}(\mathbf{w})\|^2 - \langle \nabla h(\mathbf{w}), \nabla\tilde{\Phi}(\mathbf{w})\rangle}{h(\mathbf{w})^2} \le \frac{\|\nabla\tilde{\Phi}(\mathbf{w})\|^2}{h(\mathbf{w})^2} + G'\frac{\|\nabla F(\mathbf{w})\|^2 + \|\nabla\tilde{\Phi}(\mathbf{w})\|^2}{h(\mathbf{w})^2} \le (G'+1)C' + G'\frac{\|\nabla F(\mathbf{w})\|^2}{h(\mathbf{w})^2}.$$

Recalling $h(\mathbf{w}) \ge B \ge 1$, an upper bound on the above is then simply

$$C'' := (G+1)C' + G'G_F^2, \tag{34}$$

Here C' is from (22), G_F is as per (31), and G' is as defined above. As discussed above, this bound still applies in the $\mathbf{w} \in \mathcal{S}^c$ case. Thus, we have for all $\mathbf{w}$ that

$$\frac{\|\nabla\tilde{\Phi}(\mathbf{w})\|^2 - \langle \nabla h(\mathbf{w}), \nabla\tilde{\Phi}(\mathbf{w})\rangle}{h(\mathbf{w})^2} \le C''.$$

Step b. From here, we can conclude a WPI analogously to Step b, Subsection D.1, the one difference being that we need to control the 'error term' $\int_{\mathcal{S}} f(\mathbf{w})^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\left|\int_{\mathcal{S}} f(\mathbf{w})^2 \frac{-\mathcal{L}\tilde{\Phi}(\mathbf{w})}{h(\mathbf{w})}\mathrm{d}\mu_\beta\right|$. For convenience, let

$$\mathrm{err}(f) := \int_{\mathcal{S}} f(\mathbf{w})^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\left|\int_{\mathcal{S}} f(\mathbf{w})^2 \frac{-\mathcal{L}\tilde{\Phi}(\mathbf{w})}{h(\mathbf{w})}\mathrm{d}\mu_\beta\right|. \tag{35}$$

Recalling (32) and using that $f^2 \ge 0$ thus gives

$$\begin{aligned}
&\int f(\mathbf{w})^2 \frac{\tilde{g}(F(\mathbf{w}))}{\tilde{g}(F(\mathbf{w})) + B}\mathrm{d}\mu_\beta \\
&\le \frac{1}{\beta}\int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\int C'' f(\mathbf{w})^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\int f(\mathbf{w})^2 \frac{|\Delta\tilde{\Phi}(\mathbf{w})|}{\tilde{g}(F(\mathbf{w})) + B}\mathrm{d}\mu_\beta + \mathrm{err}(f) \\
&\le \frac{1}{\beta}\int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\int C'' f(\mathbf{w})^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\int f(\mathbf{w})^2 \frac{d\tilde{L}}{\tilde{g}(F(\mathbf{w})) + \tilde{L}}\mathrm{d}\mu_\beta + \mathrm{err}(f) \\
&\le \frac{1}{\beta}\int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\int f(\mathbf{w})^2 (d + C'')\mathrm{d}\mu_\beta + \mathrm{err}(f),
\end{aligned}$$

where we used (34) and the $\tilde{L}$-smoothness of $\tilde{\Phi}$. Notice $\frac{\tilde{g}(t)}{\tilde{g}(t)+B}$ is non-decreasing as $\tilde{g}$ is non-decreasing. We thus obtain:

$$\begin{aligned}
&\int f(\mathbf{w})^2 \frac{\tilde{g}(l_b)}{\tilde{g}(l_b) + B}\mathrm{d}\mu_\beta \\
&= \int_{\mathcal{U}^c} f(\mathbf{w})^2 \frac{\tilde{g}(l_b)}{\tilde{g}(l_b) + B}\mathrm{d}\mu_\beta + \int_{\mathcal{U}} f(\mathbf{w})^2 \frac{\tilde{g}(l_b)}{\tilde{g}(l_b) + B}\mathrm{d}\mu_\beta \\
&\le \int f(\mathbf{w})^2 \frac{\tilde{g}(F(\mathbf{w}))}{\tilde{g}(F(\mathbf{w})) + B}\mathrm{d}\mu_\beta + \int_{\mathcal{U}} f(\mathbf{w})^2 \frac{\tilde{g}(l_b)}{\tilde{g}(l_b) + B}\mathrm{d}\mu_\beta
\end{aligned}$$

$$\leq \frac{1}{\beta}\int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\int f(\mathbf{w})^2\big(d + C''\big)\mathrm{d}\mu_\beta + \frac{\tilde{g}(l_b)}{\tilde{g}(l_b)+B}\int_{\mathcal{U}} f(\mathbf{w})^2 \mathrm{d}\mu_\beta + \mathrm{err}(f). \tag{36}$$

Exactly as in Subsection D.1, using Assumption 3.1 and the definition $f = \psi - \alpha$ (the choice of α is crucial), we obtain

$$\int_{\mathbf{U}} f(\mathbf{w})^2 \mathrm{d}\mu_\beta \leq \mathsf{C}_{\text{PI, LOCAL}} \int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta.$$

Applying this in (36), we obtain

$$\begin{aligned}
&\int f(\mathbf{w})^2 \frac{\tilde{g}(l_b)}{\tilde{g}(l_b)+B} \mathrm{d}\mu_\beta \\
&\leq \frac{1}{\beta}\int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\int f(\mathbf{w})^2\big(d + C''\big)\mathrm{d}\mu_\beta + \frac{\tilde{g}(l_b)}{\tilde{g}(l_b)+B}\cdot \mathsf{C}_{\text{PI, LOCAL}}\int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta + \mathrm{err}(f).
\end{aligned}$$

For $\beta \geq 2\Big(1 + \frac{B}{\tilde{g}(l_b)}\Big)(d + C'') = \Omega(d)$, this gives

$$\frac{\tilde{g}(l_b)}{2(\tilde{g}(l_b)+B)}\int f(\mathbf{w})^2 \mathrm{d}\mu_\beta \leq \frac{1}{\beta}\int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta + \frac{\tilde{g}(l_b)}{\tilde{g}(l_b)+B}\cdot \mathsf{C}_{\text{PI, LOCAL}}\int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta + \mathrm{err}(f).$$

Rearranging this inequality and converting back to ψ, recalling the definition of variance, and noting $\nabla f = \nabla\psi$ gives:

$$\begin{aligned}
\mathbb{V}_{\mu_\beta}[\psi] &\leq \int (\psi - \alpha)^2 \mathrm{d}\mu_\beta \\
&= \int f^2 \mathrm{d}\mu_\beta \\
&\leq \left(2\mathsf{C}_{\text{PI, LOCAL}} + \frac{2}{\beta}\left(1 + \frac{B}{\tilde{g}(l_b)}\right)\right)\int \|\nabla f\|^2 \mathrm{d}\mu_\beta + 2\left(1 + \frac{B}{\tilde{g}(l_b)}\right)\mathrm{err}(f) \\
&= \left(2\mathsf{C}_{\text{PI, LOCAL}} + \frac{2}{\beta}\left(1 + \frac{B}{\tilde{g}(l_b)}\right)\right)\int \|\nabla \psi\|^2 \mathrm{d}\mu_\beta + 2\left(1 + \frac{B}{\tilde{g}(l_b)}\right)\mathrm{err}(f).
\end{aligned}$$

Finally, we control the error term $\mathrm{err}(f)$. First note for $\mathbf{w} \in \mathcal{S}$, by definition of B in (30),

$$\left|\frac{-\mathcal{L}\tilde{\Phi}(\mathbf{w})}{h(\mathbf{w})}\right| \leq \frac{\beta\|\nabla F(\mathbf{w})\|\|\nabla\tilde{\Phi}(\mathbf{w})\|}{g(F(\mathbf{w}))+B} + \frac{|\Delta\tilde{\Phi}|}{h(\mathbf{w})} \leq \frac{\beta G_F G_\Phi}{G_F G_\Phi} + \frac{d\tilde{L}}{\tilde{L}} \leq \beta + d.$$

Next, recall $f = \psi - \alpha$ where $\alpha = \frac{1}{\mu_\beta(\mathcal{U})}\int_{\mathcal{U}} \psi \mathrm{d}\mu_\beta = \int_{\mathcal{U}} \psi \mathrm{d}\mu_{\beta,\text{LOCAL}}$ is defined as before. Note $\alpha \in [\inf\psi, \sup\psi]$. Note for all $\mathbf{w}$,

$$\begin{aligned}
\psi(\mathbf{w}) - \alpha &\leq \sup\psi - \inf\psi = \mathrm{osc}(\psi), \\
\psi(\mathbf{w}) - \alpha &\geq \inf\psi - \sup\psi = -\mathrm{osc}(\psi).
\end{aligned}$$

Consequently, we have for all $\mathbf{w}$,

$$f(\mathbf{w})^2 = (\psi(\mathbf{w}) - \alpha)^2 \leq \mathrm{osc}(\psi)^2.$$

Thus, recalling $\beta \geq d$, we obtain

$$\mathrm{err}(f) = \int_{\mathcal{S}} f(\mathbf{w})^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\left|\int_{\mathcal{S}} f(\mathbf{w})^2 \frac{-\mathcal{L}\tilde{\Phi}(\mathbf{w})}{h(\mathbf{w})}\mathrm{d}\mu_\beta\right|$$

$$\leq \operatorname{osc}(\psi)^2 \mu_\beta(\mathcal{S})\Big(1 + \frac{1}{\beta}(d+\beta)\Big) \leq 3\operatorname{osc}(\psi)^2 \mu_\beta(\mathcal{S}).$$

Consequently we have

$$\mathbb{V}_{\mu_\beta}[\psi] \leq \left(2\mathsf{C}_{\text{PI, LOCAL}} + \frac{2}{\beta}\left(1 + \frac{\tilde{L}}{\tilde{g}(l_b)}\right)\right) \int \|\nabla \psi\|^2 \mathrm{d}\mu_\beta + 6\left(1 + \frac{B}{\tilde{g}(l_b)}\right)\mu_\beta(\mathcal{S})\operatorname{osc}(\psi)^2.$$

Recalling ψ is an arbitrary test function, this shows that μ_β satisfies a WPI of the form

$$\left(2\mathsf{C}_{\text{PI, LOCAL}} + \frac{2}{\beta}\left(1 + \frac{B}{\tilde{g}(l_b)}\right), 6\left(1 + \frac{B}{\tilde{g}(l_b)}\right)\mu_\beta(\mathcal{S})\right) \text{ for } \beta \geq 2\left(1 + \frac{B}{\tilde{g}(l_b)}\right)(d + C''), \tag{37}$$

where B is defined in (30) and C'' is defined in (34). ■

Remark 33 *Notice that in the region $\mathcal{S}$ where GF/GD do not work, one would generally expect $\|\nabla F(\mathbf{w})\|$ and thus G_F to be very small. Moreover, the dependence on F-dependent constants above can be optimized in the above analysis; we made little effort to do so.*

Remark 34 *Note that the construction of $\tilde{\Phi}$ is unnecessary if Φ is smooth, and in this case the expressions simplify analogously to Remark 30. However, in this setting, we cannot assume $\mathcal{S} \subseteq \mathbb{B}(\mathbf{w}^\star, R+1)$.*

Proof of Corollary 17. If we only have a $(\mathsf{C}_{\text{WPI, LOCAL}}, \delta_{\text{LOCAL}})$-WPI for $\mu_{\beta,\text{LOCAL}}$ rather than Assumption 3.1, we can proceed as follows to prove a WPI for μ_β. Perform the exact same moves as in Subsection D.1 up until establishing (24), including our choice of arbitrary test function ψ and f defined in terms of ψ, none of which utilize Assumption 3.1. Follow the exact same notation as in that proof. These same exact steps again give (24):

$$\int f^2 \frac{\tilde{g}(l_b)}{\tilde{g}(l_b) + \tilde{L}} \mathrm{d}\mu_\beta \leq \frac{1}{\beta}\int \|\nabla f\|^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\int f^2\big(d + C'\big)\mathrm{d}\mu_\beta + \frac{\tilde{g}(l_b)}{\tilde{g}(l_b) + \tilde{L}}\int_{\mathcal{U}} f^2 \mathrm{d}\mu_\beta. \tag{38}$$

Now rather than utilizing a PI for $\mu_{\beta,\text{LOCAL}}$ which we do not have, use the $(\mathsf{C}_{\text{WPI, LOCAL}}, \delta_{\text{LOCAL}})$-WPI for $\mu_{\beta,\text{LOCAL}}$ on the test function f to obtain

$$\mathbb{V}_{\mu_{\beta,\text{LOCAL}}}(f) \leq \mathsf{C}_{\text{WPI, LOCAL}} \int \|\nabla f\|^2 \mathrm{d}\mu_{\beta,\text{LOCAL}} + \delta_{\text{LOCAL}}\operatorname{osc}(f)^2.$$

The left hand side above also equals

$$\int f^2 \mathrm{d}\mu_{\beta,\text{LOCAL}} - \left(\int f \mathrm{d}\mu_{\beta,\text{LOCAL}}\right)^2 = \frac{1}{\mu_\beta(\mathcal{U})}\int_{\mathcal{U}} f^2 \mathrm{d}\mu_\beta - \frac{1}{\mu_\beta(\mathcal{U})^2}\left(\int_{\mathcal{U}} f \mathrm{d}\mu_\beta\right)^2.$$

That is, we have

$$\frac{1}{\mu_\beta(\mathcal{U})}\int_{\mathcal{U}} f^2 \mathrm{d}\mu_\beta - \frac{1}{\mu_\beta(\mathcal{U})^2}\left(\int_{\mathcal{U}} f \mathrm{d}\mu_\beta\right)^2 \leq \frac{\mathsf{C}_{\text{WPI, LOCAL}}}{\mu_\beta(\mathcal{U})}\int_{\mathcal{U}} \|\nabla f\|^2 \mathrm{d}\mu_\beta + \delta_{\text{LOCAL}}\operatorname{osc}(f)^2.$$

Recalling the definition of f in terms of ψ, the above rearranges to

$$\int_{\mathcal{U}} f^2 \mathrm{d}\mu_\beta \leq \mathsf{C}_{\text{WPI, LOCAL}}\int_{\mathcal{U}} \|\nabla f\|^2 \mathrm{d}\mu_\beta + \mu_\beta(\mathcal{U}) \cdot \delta_{\text{LOCAL}}\operatorname{osc}(f)^2$$
$$+ \frac{1}{\mu_\beta(\mathcal{U})}\left(\int_{\mathcal{U}}\left(\psi(\mathbf{w}) - \frac{1}{\mu_\beta(\mathcal{U})}\int_{\mathcal{U}} \psi(\mathbf{w})\mathrm{d}\mu_\beta\right)\mathrm{d}\mu_\beta\right)^2$$

$$\leq \mathsf{C}_{\text{WPI, LOCAL}} \int \|\nabla f\|^2 \mathrm{d}\mu_\beta + \delta_{\text{LOCAL}} \operatorname{osc}(f)^2.$$

Applying this in (38), we obtain

$$\int f^2 \frac{\tilde{g}(l_b)}{\tilde{g}(l_b) + \tilde{L}} \mathrm{d}\mu_\beta \leq \frac{1}{\beta} \int \|\nabla f\|^2 \mathrm{d}\mu_\beta + \frac{1}{\beta} \int f^2 (d + C') \mathrm{d}\mu_\beta$$
$$+ \frac{\tilde{g}(l_b)}{\tilde{g}(l_b) + \tilde{L}} \left(\mathsf{C}_{\text{WPI, LOCAL}} \int \|\nabla f\|^2 \mathrm{d}\mu_\beta + \delta_{\text{LOCAL}} \operatorname{osc}(f)^2 \right).$$

If $\beta \geq 2\left(1 + \frac{\tilde{L}}{\tilde{g}(l_b)}\right)(d + C') = \Omega(d)$, this gives

$$\frac{\tilde{g}(l_b)}{2(\tilde{g}(l_b) + \tilde{L})} \int f^2 \mathrm{d}\mu_\beta \leq \frac{1}{\beta} \int \|\nabla f\|^2 \mathrm{d}\mu_\beta + \frac{\tilde{g}(l_b)}{\tilde{g}(l_b) + \tilde{L}} \left(\mathsf{C}_{\text{WPI, LOCAL}} \int \|\nabla f\|^2 \mathrm{d}\mu_\beta + \delta_{\text{LOCAL}} \operatorname{osc}(f)^2 \right).$$

Now, we rearrange above the inequality and convert back to ψ. Recalling the definition of variance and noting $\nabla f = \nabla \psi$, and noting ψ is just a constant shift of f and hence $\operatorname{osc}(\psi) = \operatorname{osc}(f)$, we obtain:

$$\begin{aligned}
\mathbb{V}_{\mu_\beta}[\psi] &\leq \int (\psi - \alpha)^2 \mathrm{d}\mu_\beta \\
&= \int f^2 \mathrm{d}\mu_\beta \\
&\leq \left(2\mathsf{C}_{\text{WPI, LOCAL}} + \frac{2}{\beta}\left(1 + \frac{\tilde{L}}{\tilde{g}(l_b)}\right)\right) \int \|\nabla f\|^2 \mathrm{d}\mu_\beta + 2\delta_{\text{LOCAL}} \operatorname{osc}(f)^2 \\
&= \left(2\mathsf{C}_{\text{WPI, LOCAL}} + \frac{2}{\beta}\left(1 + \frac{\tilde{L}}{\tilde{g}(l_b)}\right)\right) \int \|\nabla \psi\|^2 \mathrm{d}\mu_\beta + 2\delta_{\text{LOCAL}} \operatorname{osc}(\psi)^2.
\end{aligned}$$

Recalling ψ is an arbitrary test function, this shows that μ_β satisfies a WPI with constants

$$\left(2\mathsf{C}_{\text{WPI, LOCAL}} + \frac{2}{\beta}\left(1 + \frac{\tilde{L}}{\tilde{g}(l_b)}\right), 2\delta_{\text{LOCAL}} \right) \text{ for } \beta \geq 2\left(1 + \frac{\tilde{L}}{\tilde{g}(l_b)}\right)(d + C'). \tag{39}$$

Again, $\tilde{L}$ comes from (19), C' comes from (22).

The extension to the setting of Theorem 15 follows the exact same steps, and shows that μ_β satisfies a Weak Poincaré Inequality of the form

$$\left(2\mathsf{C}_{\text{WPI, LOCAL}} + \frac{2}{\beta}\left(1 + \frac{B}{\tilde{g}(l_b)}\right), 6\left(1 + \frac{B}{\tilde{g}(l_b)}\right)\mu_\beta(\mathcal{S}) + 2\delta_{\text{LOCAL}} \right) \text{ for } \beta \geq 2\left(1 + \frac{B}{\tilde{g}(l_b)}\right)(d + C''),$$

where again B is defined in (30) and C'' is defined in (34). ∎

D.3. Proofs of Corollary 19, Corollary 20

Proof of Corollary 19. First, apply Theorem 12 to obtain

$$\mathsf{C}_{\text{PI}} = O(\mathsf{C}_{\text{PI, LOCAL}} + 1/\beta).$$

- Now, the first part on sampling via LMC under Assumption 3.3 follows directly as a corollary of Theorem 7 of Chewi et al. (2024) on sampling from targets satisfying a PI, which we apply with βL in place of L there as our potential in question is βF, and with Rényi divergence of order $q = 1$ (hence we obtain a result in KL) and LOI inequality of order $\alpha = 1$. The implementation for the step size is *exactly* the same as in these theorems and the corresponding implementation in Chewi et al. (2024). In particular the step size h is given by 6.10 of Chewi et al. (2024); the only change is changing L to βL exactly as mentioned above, and applying the new bounds for initialization in this setting now from Lemma 43. We appeal to Lemma 43 to control the initialization, $\mathsf{KL}(\pi_0 \| \mu_\beta)$, and the Rényi Divergence of order 2, which is $\ln(\chi^2(\pi_0 \| \mu_\beta) + 1)$. This justifies that the explicit β, d dependence of the initialization is $\tilde{O}(\beta)$ for $\beta = \Omega(d)$ up to log factors (see more discussion in Remark 44). Thus, as a direct corollary of Theorem 7 of Chewi et al. (2024), we see that LMC satisfies the following guarantee:

$$\mathsf{KL}\big(\pi_T \| \mu_\beta\big) \leq \varepsilon \text{ after } T = \tilde{O}\left(d\left(\mathsf{C}_{\text{PI, LOCAL}} + \frac{1}{\beta}\right)^{1+\frac{1}{s}} \beta^{1+\frac{3}{s}} \varepsilon^{-\frac{1}{s}} \cdot \max\left\{1, \frac{\beta^{s/2}}{d}\right\}\right) \text{ iterations.}$$

Applying Pinkser's Inequality yields the desired.

The term $\max\left\{1, \frac{\beta^{s/2}}{d}\right\}$ warrants some discussion. It arises here in the maximum of Theorem 7, Chewi et al. (2024). The second term there does not dominate, and it seems reasonable that the third term there does not dominate, as we justify in Remark 44. However, now the fourth term in the maximum could dominate, and we argue in Lemma 43 that we can take it to be $\tilde{O}(\beta)$. This gives the factor $\max\left\{1, \frac{\beta^{s/2}}{d}\right\}$.

For more details on the implementation of γ here, here $\gamma \leq \frac{1}{768Th} \leq 1$ as per Proposition 29, Chewi et al. (2024). Since $\gamma \leq 1$, applying Lemma 43 gives the claimed bounds on the initialization. T is the iteration count reported above, and the step size h is given by 6.10 of Chewi et al. (2024), with the only explicit change of changing L to βL and using the new bounds on initialization.

- The second part on sampling under the Proximal Sampler follows directly from Theorem 5.4, Altschuler and Chewi (2024), on sampling from targets satisfying a PI. The implementation for the step size is *exactly* the same as in these theorems and the corresponding implementation in Altschuler and Chewi (2024), where we take the smoothness constant in their result equal to βL, the smoothness constant of our potential βF. Here we can initialize π_0 as in Corollary 19 for *any* $\gamma \leq 1$, and simply use the first part of Lemma 43 to argue the initial divergence $\ln(\chi^2(\pi_0 \| \mu_\beta))$ is controlled by $\tilde{O}(\beta + d)$ (again see more discussion in Remark 44).

This completes the proof. ■

Note that the above is simply a corollary of our main results, and is *not* the focus of our work.

Remark 35 *Notice we only used the PI from Theorem 12 above. Indeed, there is little gain in using the LSI vs PI from Theorem 12 in the proof above. This is certainly not because is no gain in an LSI; rather, it is because our LSI bound loses about a factor of βS for $\beta = \Omega(d)$, and so combining Theorem 12 with pre-existing results on sampling under LSI does not give better results.*

Proof of Corollary 20. We first show that Assumption A.1 implies the following assumption from Lytras and Mertikopoulos (2024), allowing us to use their results:

Assumption D.1 (Assumption 1 from Lytras and Mertikopoulos (2024)) *Suppose F satisfies the following properties, from Assumption 1, Lytras and Mertikopoulos (2024):*

- *Polynomial Lipschitz Continuity: for some $s_1, L_1' > 0$, we have for all $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d$,*

$$\|\nabla F(\mathbf{w}_1) - \nabla F(\mathbf{w}_2)\| \leq L_1'(1 + \|\mathbf{w}_1\| + \|\mathbf{w}_2\|)^{s_1} \|\mathbf{w}_1 - \mathbf{w}_2\|.$$

- *Weak Dissipativity: for some $s_2 \geq 1$, $A_2, b_2 > 0$, we have for all $\mathbf{w} \in \mathbb{R}^d$,*

$$\langle \nabla F(\mathbf{w}), \mathbf{w} \rangle \geq A_2 \|\mathbf{w}\|^{s_2} - b_2.$$

- *Polynomial Jacobian Growth: for some $L_3, s_3 > 0$ and all $k \geq 2$ for which the following is well-defined, we have for all $\mathbf{w} \in \mathbb{R}^d$,*

$$\max\Big(\|\nabla F(\mathbf{w})\|, \big\|\nabla^k F(\mathbf{w})\big\|_{\mathrm{op}} \Big) \leq L_3 (1 + \|\mathbf{w}\|)^{2s_3}.$$

To verify this, take $k = 2$ in Assumption A.1, and note for any $\mathbf{w} = t\mathbf{w}_1 + (1-t)\mathbf{w}_2$ for $0 \leq t \leq 1$ that

$$\big\|\nabla^2 F(\mathbf{w})\big\| \leq L_3 (1 + \|t\mathbf{w}_1 + (1-t)\mathbf{w}_2\|)^{2s_3} \leq L_3 (1 + \|\mathbf{w}_1\| + \|\mathbf{w}_2\|)^{2s_3}.$$

Consequently as this holds for all $\mathbf{w}$ in the line segment $\overline{\mathbf{w}_1\mathbf{w}_2}$, we obtain

$$\|\nabla F(\mathbf{w}_1) - \nabla F(\mathbf{w}_2)\| \leq L_3 (1 + \|\mathbf{w}_1\| + \|\mathbf{w}_2\|)^{2s_3} \|\mathbf{w}_1 - \mathbf{w}_2\|,$$

and so from Assumption A.1, we have Assumption D.1 with $L_1' = L_3$, $s_1 = 2s_3$.

Now to establish Corollary 20, we directly apply Theorems 2 and 3 of Lytras and Mertikopoulos (2024). These results show that their relevant algorithm can yield a distribution π_T with $\mathsf{KL}\big(\pi_T \| \mu_\beta\big) \leq \varepsilon$ for large enough T. In particular:

- Theorem 2 of Lytras and Mertikopoulos (2024) shows that under Assumption D.1, if μ_β satisfies a Log-Sobolev Inequality with constant $\mathsf{C}_{\mathrm{LSI}}$, then via their algorithm wd-TULA we have

$$\mathsf{KL}\big(\pi_T \| \mu_\beta\big) \leq \varepsilon \text{ within } T = \tilde{O}\left(\frac{\mathrm{poly}(d, \beta)\mathsf{C}_{\mathrm{LSI}}}{\varepsilon} \log\left(\frac{\mathsf{KL}\big(\pi_0 \| \mu_\beta\big)}{\varepsilon} \right) \right) \text{ iterations.}$$

- Theorem 3 of Lytras and Mertikopoulos (2024) shows that under Assumption D.1, if μ_β satisfies a Poincaré Inequality with constant C_{PI}, then via their algorithm reg-TULA we can take

$$\mathsf{KL}\big(\pi_T \| \mu_\beta\big) \leq \varepsilon \text{ within } T = \tilde{O}\left(\mathrm{poly}\left(d, \beta, \mathsf{C}_{\mathrm{PI}}, \frac{1}{\varepsilon}\right) \log\left(\frac{\mathsf{KL}\big(\pi_0 \| \hat{\mu}_\beta\big)}{\varepsilon} \right) \right) \text{ iterations.}$$

  Here, $\hat{\mu}_\beta$ corresponds to $e^{-\left(\beta F(\mathbf{w}) + \eta \|\mathbf{w}\|^{2r+2}\right)}/Z$, where r is taken large enough in terms of the exponents s_1, s_2, s_3 from Assumption D.1. The degree of these polynomials also depends on s_1, s_2, s_3.

Note Assumption 1 of Lytras and Mertikopoulos (2024) is phrased in terms of the true potential βF rather than F. Their results have polynomial d dependence, but to convert these results to our setting where $\beta = \Omega(d)$, we need to track their proofs and find the explicit dependency on their parameters A, L, L', b, which are scaled up by β for us.

We explicitly make this conversion here for the reader's convenience: converting to their notation we have

$$L' = \beta L_1' = \beta L_3, A = \beta A_2, b = \beta b_2, L = \beta L_3.$$

The powers do not change: converting to their notation we still have $l' = s_1$, $a = s_2$, $l = s_3$. For the rest of this discussion, we follow the notation of Lytras and Mertikopoulos (2024) so the reader can easily reference their work.

We find that this dependency is polynomial in their guarantees from Theorems 2 and 3. In particular, we carefully track this for $\hat{C}$ from their Theorem 2 and their $\hat{C}, \dot{c}$ from their Theorem 3, and see the dependencies on these is polynomial with respect to d, A', K, L, L', b from their Assumption 1. By consequence the dependence on β is also polynomial.[19] However such dependence on problem-dependent A', K, L, L', b is not made as explicit in Lytras and Mertikopoulos (2024), and so we explicitly track this here. For more details:

- Consider their Theorem 2. The convergence rate there is given in terms of $\mathsf{C}_{\text{LSI}}, \mathsf{KL}(\pi_0 \| \mu_\beta)$, and $\hat{C}$. $\hat{C}$ bounds the discretization error, and through the proof of Lemma A.5, $\hat{C}$ is in turn given by a polynomial function of $C_{1,p}, C_p$ for integers $p \geq 0$ from their Lemmas A.3 and A.4. These quantities control various moment bounds. In turn, these are all given in terms of the C_p from their Lemma A.3 and polynomial factors in A', L, L', d (recall A', L, L' are β times our smoothness constants). C_p here is at most $(\ln C_\mu)^{2p}$ where C_μ is defined in their Lemma A.2 and controls the growth of particular exponential moments. Tracking the proof of Lemma A.2, we can see that $C_\mu \leq \exp\{\text{poly}(A, L, L', b, d)\}$. Thus $C_p \leq \text{poly}(A, L, L', b, d)$, and so $\hat{C} \leq \text{poly}(A, L, L', b, d)$.

- Consider their Theorem 3. This is derived from their Theorem 7, where the convergence rate there is given in terms of $\hat{C}$, which again controls discretization error, and $\dot{c}$, which governs the Log-Sobolev constant of a particular regularized version of the potential βF. The regularization is in particular given by $\beta F(\mathbf{w}) + \lambda \|\mathbf{w}\|^{2r+2}$. Here λ denotes the step size and we can without loss of generality take $\lambda \leq 1$.

  First we consider $\hat{C}$. Analyzing the proof of Theorem 7, we see that it is given by the sum of $C_{\text{tam}}^{\text{reg}}$ and $C_{\text{onestep}}^{\text{reg}}$ from Lemmas C.2, C.3. In turn, these quantities are controlled exactly the same way by the moment bounds as in Lemmas A.5, and in turn Lemmas A.3 and A.4, except now we are dealing with the regularized potential $\beta F + \lambda \|\mathbf{w}\|^{2r+2}$ rather than the original potential βF (this is shown for example in their Lemma C.6). As noted in the article, we can prove analogous moment bounds the same way, with still dependence that is $\text{poly}(A, L, L', b, d)$. This is because the proof of their Lemma A.6 shows the regularized potential still satisfies their Assumption 1 parts A1 and A2 and a result analogous to Lemma A.1, with smoothness parameters only a universal constant shift from A, L, L', b for regularization $\lambda \leq 1$. These are all the conditions needed to prove Lemma A.2, which in turn give the desired bounds Lemma A.3 and A.4, for the regularized potential.

  Next we consider $\dot{c}$. The dependence of $\dot{c}$ on λ is given in Proposition 3.8, Lytras and Mertikopoulos (2024), which upon converting to our notation, is $\left(\frac{1}{\lambda}\right)^{\frac{1}{r+1} + \frac{s_1}{2r - s_1}}$. We need $\frac{1}{r+1} + \frac{s_1}{2r - s_1} \leq 1$ to obtain a meaningful convergence rate, and indeed we can make $\frac{1}{r+1} + \frac{s_1}{2r - s_1} \leq \frac{1}{2}$ by taking r large enough in terms of s_1. The dependence of $\dot{c}$ on all other parameters is given from their equation C.8 in the proof of their Proposition A.4 (we note that the third term in that equation is a typo and should read, following their notation, $\frac{K_\lambda}{A_{\text{reg}}}$ from using Theorem 3.15 of Menz and Schlichting (2014)). We can check that, by what we have argued on moment control in the above paragraph, all the other parameters $A_{\text{reg}}, K_\lambda, \pi_{\text{reg}}(\|x\|^2)$ and Poincaré constant of the Gibbs measure of the regularized potential all depend polynomially on A', L, L', b, d. Hence $\dot{c}$ depends polynomially on A', L, L', b, d.

  We conclude upon applying the same rationale as Theorem 7 and Corollary 4 of Lytras and Mertikopoulos (2024).

One additional point of consideration is these results contain dependence on initial divergences

$$\mathsf{KL}(\pi_0 \| \mu_\beta), \mathsf{KL}(\pi_0 \| \hat{\mu}_\beta).$$

19. This is to be expected; in many results on discrete-time LMC, e.g. Chewi et al. (2024), dependence on smoothness constants (which are also scaled up by β here) are polynomial.

We argue that these both can be controlled by $\tilde{O}(d\beta)$ in Lemma 46, given appropriate initialization. As noted on footnote 1 of page 7 of Lytras and Mertikopoulos (2024), or just by tracking their proof, we note that their result holds for any initialization (at the expense of a different price for initialization $\mathsf{KL}(\pi_0\|\mu_\beta), \mathsf{KL}(\pi_0\|\hat{\mu}_\beta)$). Note since these initializations are polynomial in d, β, they do not affect the claimed rate or Corollary 20 (as they appear in the logarithm, as per Lytras and Mertikopoulos (2024)). Putting all this together, combining with Points 1 and 2, and using Pinkser's Inequality gives Corollary 20. ∎

We emphasize that we just cite the result of Lytras and Mertikopoulos (2024) and made no attempt to optimize the polynomial dependency. The focus on our work is on proving isoperimetric inequalities. Moreover, while the dependence indicated above is polynomial, again note the degree of the polynomials in question depends on the exponents s_1, s_2, s_3 from Assumption D.1.

D.4. Proofs of Subsection A.2

We first verify that $\hat{\mu}_\beta$, μ_β are indeed close in TV distance:

Lemma 36 *Defining δ as in Corollary 22, we have* $\mathsf{TV}(\hat{\mu}_\beta, \mu_\beta) \le 3\delta$.

Proof. Let $I = \int_{\mathbb{B}(\mathbf{w},R-1)} e^{-\beta F(\mathbf{w})}\mathrm{d}\mathbf{w}$. By construction of $\hat{F}$, we also have $I = \int_{\mathbb{B}(\mathbf{w},R-1)} e^{-\beta \hat{F}(\mathbf{w})}\mathrm{d}\mathbf{w}$. Let $I_1 = \int_{\mathbb{B}(\mathbf{w},R-1)^c} e^{-\beta F(\mathbf{w})}\mathrm{d}\mathbf{w}$, $I_2 = \int_{\mathbb{B}(\mathbf{w},R-1)^c} e^{-\beta \hat{F}(\mathbf{w})}\mathrm{d}\mathbf{w}$. Note $I_2 \le I_1$ as $\hat{F} \ge F$ on $\mathbb{B}(\mathbf{w}, R-1)^c$. Consequently, recalling the definition of δ, we have

$$1 \ge \frac{I}{I+I_2} \ge \frac{I}{I+I_1} \ge 1-\delta, \text{ thus } 0 \le \frac{I_1}{I+I_1}, \frac{I_2}{I+I_2} \le \delta.$$

Now consider any subset $\mathcal{A} \subseteq \mathbb{R}^d$, and let $\mathcal{A}_1 = \mathcal{A} \cap \mathbb{B}(\mathbf{w}, R-1)$, $\mathcal{A}_2 = \mathcal{A} \cap \mathbb{B}(\mathbf{w}, R-1)^c$. Note $F, \hat{F}$ agree on $\mathcal{A}_1$ and so $\int_{\mathcal{A}_1} e^{-\beta F(\mathbf{w})}\mathrm{d}\mathbf{w} = \int_{\mathcal{A}_1} e^{-\beta \hat{F}(\mathbf{w})}\mathrm{d}\mathbf{w} = xI$ for some $x \in [0,1]$. Let $Y_1 = \int_{\mathcal{A}_2} e^{-\beta F(\mathbf{w})}\mathrm{d}\mathbf{w}$, let $Y_2 = \int_{\mathcal{A}_2^c} e^{-\beta \hat{F}(\mathbf{w})}\mathrm{d}\mathbf{w}$, and note $Y_1 \le I_1$, $Y_2 \le I_2$. Thus we obtain

$$\begin{aligned}
\left|\hat{\mu}_\beta(\mathcal{A}) - \mu_\beta(\mathcal{A})\right| &= \left|\frac{xI}{I+I_1} - \frac{xI}{I+I_2} + \frac{Y_1}{I+I_1} - \frac{Y_2}{I+I_2}\right| \\
&\le \left|\frac{xI}{I+I_1} - \frac{xI}{I+I_2}\right| + \left|\frac{Y_1}{I+I_1} - \frac{Y_2}{I+I_2}\right| \\
&\le x\left|\frac{I}{I+I_1} - \frac{I}{I+I_2}\right| + \frac{Y_1}{I+I_1} + \frac{Y_2}{I+I_2} \\
&\le \delta + \delta + \delta = 3\delta.
\end{aligned}$$

This applies for all $\mathcal{A} \subset \mathbb{R}^d$, and we conclude. ∎

Proof of Proposition 21. Let $\mathcal{U} = \mathbb{B}(\mathcal{W}^\star, r(l_b))$ for any l_b satisfying Assumption 3.1.

Part 1: Modifying the Interpolation Argument Recall for a suitable bump function $\chi_F \in [0,1]$ which we will define later, we defined

$$\tilde{F}(\mathbf{w}) := \begin{cases} F(\mathbf{w}) & : \|\mathbf{w}-\mathbf{w}^\star\| \le R-1 \\ F(\mathbf{w}) + \chi_F(\mathbf{w}) \cdot \lambda_{\text{REG}}(\|\mathbf{w}-\mathbf{w}^\star\|^2+1) & : R-1 < \|\mathbf{w}-\mathbf{w}^\star\| < R \\ F(\mathbf{w}) + \lambda_{\text{REG}}(\|\mathbf{w}-\mathbf{w}^\star\|^2+1) & : R \le \|\mathbf{w}-\mathbf{w}^\star\|, \end{cases}$$

where $\lambda_{\text{REG}} = L$.[20] Also let

$$L_{b,1} = \inf_{R-1 \le \|\mathbf{w}-\mathbf{w}^\star\| \le R} F(\mathbf{w}).$$

20. In fact, λ_{REG} can be any upper bound on L, which can be seen by tracking the following proof.

By assumption that $\mathbb{B}(\mathcal{W}^\star, r(l_b)) \subseteq \mathbb{B}(\mathbf{w}^\star, R-1)$, we have $L_{b,1} \geq l_b$.

We now show that we can perform the same interpolation steps as in the proof of Theorem 12 in Subsection D.1, Step 1, to create $\tilde{\Phi}$, except using $\tilde{F}$ in place of F. From here, very similar steps as the proof of Theorem 12 in Subsection D.1 prove that $\hat{\mu}_\beta \propto \exp(-\beta \hat{F})$ satisfies a PI. To this end, define the interpolators as follows. First define

$$M = \left\{ \sup_{\mathbf{w}\in\mathbb{B}(\mathbf{w}^\star,R)} \Phi(\mathbf{w}) + \sup_{\mathbf{w}\in\mathbb{B}(\mathbf{w}^\star,R)} F(\mathbf{w}) \right\} \vee \frac{1}{\lambda}\left(\frac{1}{4}g(L_{b,1}) + 1\right). \tag{40}$$

Now let $\chi(\mathbf{w}) = p(\|\mathbf{w}-\mathbf{w}^\star\| - (R-1))$ be the interpolator from the proof in Subsection D.1, where $p(x) = \frac{e^{-1/x^2}}{e^{-1/x^2}+e^{-1/(1-x^2)}}$. Recall the derivatives of p, and hence $\|\nabla\chi(\mathbf{w})\|, \|\nabla^2\chi(\mathbf{w})\|_{\text{op}}$, are upper bounded by B for a universal (F-independent) constant B, and that p is differentiable to all orders. As per Lemma 48, we know p is increasing on $[0,1]$ as well. (We extend p to $[0,1]$ by $p(0) = 0, p(1) = 1$, which clearly preserves all these properties.)

Let σ_Φ be a bijection from $[0,1]$ to itself such that $p(\sigma_\Phi(1/2)) = 1/2$. Clearly we can choose σ_Φ to be infinitely differentiable, increasing, and with first and second derivatives bounded by a universal, F-independent constant. Now define define the interpolator χ_Φ for Φ by

$$p_\Phi = p \circ \sigma_\Phi, \chi_\Phi(\mathbf{w}) = p_\Phi(\|\mathbf{w}-\mathbf{w}^\star\| - (R-1)).$$

Consequently, $\chi_\Phi(1/2) = 1/2$, χ_Φ is increasing, and χ_Φ has gradient norm and Hessian operator norm bounded by a universal constant B_Φ.

Next let

$$c_F := \frac{g(L_{b,1})}{8\lambda_{\text{REG}}(R^2+1)\rho_\Phi(M)}, t_{\text{THRES},F} = 1/2.$$

Let σ_F be a bijection from $[0,1]$ to itself such that $p(\sigma_\Phi(1/2)) = c_F$. Clearly we can choose σ_F to be infinitely differentiable, increasing, and with first and second derivatives bounded by a *c_F-dependent* constant, which in turn depends on F, Φ in turn. Let $\tilde{\chi}_F$ be defined by

$$q_F = p \circ \sigma_F, \tilde{\chi}_F(\mathbf{w}) = q_F(\|\mathbf{w}-\mathbf{w}^\star\| - (R-1)).$$

Hence q_F is increasing and $q_F(1/2) = c_F$. Now define the interpolator χ_F for F by

$$\chi_F(\mathbf{w}) = \int_0^{\|\mathbf{w}-\mathbf{w}^\star\|-(R-1)} q_F(t)\mathrm{d}t. \tag{41}$$

It follows that χ_F is increasing. Also define $p_F(x) = \int_0^x q_F(t)\mathrm{d}t$. Thus $p_F' = q_F$ and p_F' is increasing, $p_F'(1/2) = c_F$, and that

$$\chi_F(\mathbf{w}) = p_F(\|\mathbf{w}-\mathbf{w}^\star\| - (R-1)).$$

Also, notice for $\|\mathbf{w}-\mathbf{w}^\star\| - (R-1) \leq t_{\text{THRES},F}$,

$$\chi_F(\mathbf{w}) = \int_0^{\|\mathbf{w}-\mathbf{w}^\star\|-(R-1)} p_F'(t)\mathrm{d}t \leq \sup_{0\leq t\leq\|\mathbf{w}-\mathbf{w}^\star\|-(R-1)} p_F'(t), \text{ thus } p_F(t) \leq c_F \text{ for } t \leq 1/2. \tag{42}$$

It also follows by the above discussion that χ_F has gradient norm and Hessian operator norm bounded by an F-dependent parameter B_F.

Finally, let

$$\Phi_2 = c_{\text{WGT}}\|\mathbf{w}-\mathbf{w}^\star\|^2 + 2M$$

where c_{WGT} is defined by

$$c_{\text{WGT}} = \frac{g(L_{b,1})}{\lambda_{\text{REG}}(R-1)((R-1)^2+1)c_F} \vee \frac{2\rho_\Phi(M)R}{(R-1)^2}.$$

This defines how much we regularize by $\|\mathbf{w}-\mathbf{w}^\star\|^2$ to ensure this construction is successful. Notice $\Phi_2 \geq \Phi$ on $\mathbb{B}(\mathbf{w}^\star, R)$. In terms of Φ_2, define

$$\tilde{\Phi}(\mathbf{w}) := \chi_\Phi(\mathbf{w})\Phi_2(\mathbf{w}) + (1-\chi_\Phi(\mathbf{w}))\Phi(\mathbf{w}). \tag{43}$$

We first show:

Lemma 37 *$\hat{F}$ is smooth with smoothness constant $O(1)$ (here $O(\cdot)$ hides problem-dependent parameters).*

Proof. This is evident for $\|\mathbf{w}-\mathbf{w}^\star\| \leq R-1$, $\|\mathbf{w}-\mathbf{w}^\star\| \geq R$, where it is straightforward to verify that $\|\nabla^2\tilde{F}\| \leq 3L$. Otherwise, we have

$$\nabla\tilde{F} = \nabla F + \nabla\chi_F \cdot \lambda_{\text{REG}}(\|\mathbf{w}-\mathbf{w}^\star\|^2+1) + \chi_F \cdot 2\lambda_{\text{REG}}(\mathbf{w}-\mathbf{w}^\star),$$

and

$$\begin{aligned}\nabla^2\tilde{F} = \nabla^2 F + \nabla^2\chi_F \cdot \lambda_{\text{REG}}(\|\mathbf{w}-\mathbf{w}^\star\|^2+1) &+ \nabla\chi_F \cdot 2\lambda_{\text{REG}}(\mathbf{w}-\mathbf{w}^\star)^T \\ &+ \nabla\chi_F \cdot 2\lambda_{\text{REG}}(\mathbf{w}-\mathbf{w}^\star) + 2\lambda_{\text{REG}}\chi_F.\end{aligned}$$

Recalling $\lambda_{\text{REG}} = L$, Triangle Inequality thus gives

$$\|\nabla^2\tilde{F}\| \leq L + LB_F(R^2+1) + 4LB_F R + 2LB_F.$$

This proves this Lemma. ∎

The benefit of regularizing is shown via the following Lemma.

Lemma 38 *For $\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R)^c$, we have*

$$\left\langle\nabla\Phi_2(\mathbf{w}), \nabla\tilde{F}(\mathbf{w})\right\rangle \geq g(F(\mathbf{w})).$$

Proof of Lemma 38. For such $\mathbf{w}$,

$$\begin{aligned}\langle\nabla\Phi_2(\mathbf{w}), \nabla F(\mathbf{w})\rangle &= \langle\nabla\Phi_2(\mathbf{w}), \nabla F(\mathbf{w}) + 2\lambda_{\text{REG}}(\mathbf{w}-\mathbf{w}^\star)\rangle \\ &= 2c_{\text{WGT}}\Big(\langle\mathbf{w}-\mathbf{w}^\star, \nabla F(\mathbf{w})\rangle + 2\lambda_{\text{REG}}\|\mathbf{w}-\mathbf{w}^\star\|^2\Big).\end{aligned}$$

Thus we have

$$\langle\nabla\Phi_2(\mathbf{w}), \nabla F(\mathbf{w})\rangle \geq 2c_{\text{WGT}}\Big(2\lambda_{\text{REG}}\|\mathbf{w}-\mathbf{w}^\star\|^2 - L\|\mathbf{w}-\mathbf{w}^\star\|\Big) \geq L\|\mathbf{w}-\mathbf{w}^\star\|^2 \geq g(F(\mathbf{w})),$$

where the last inequality follows from L-smoothness of F and that $g(x) = \lambda x$ for $\lambda \leq 1$. ∎

From Lemma 38 and the definition of $\Phi_2(\cdot)$, we directly obtain the following Corollary.

Corollary 39 *For $\mathbf{w}$ with $\|\mathbf{w}-\mathbf{w}^\star\| \in [R-1, R]$, we have*

$$\left\langle\mathbf{w}-\mathbf{w}^\star, \nabla\tilde{F}(\mathbf{w})\right\rangle \geq 0.$$

Now we break into cases and show that $\tilde{\Phi}$ is still a valid Lyapunov function, in an appropriate sense:

- For $\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R-1)$, as $\chi_\Phi, \chi_F \equiv 0$ holds identically in this set, we have

$$\left\langle \nabla\tilde{\Phi}(\mathbf{w}), \nabla\tilde{F}(\mathbf{w})\right\rangle \equiv \langle \nabla\Phi(\mathbf{w}), \nabla F(\mathbf{w})\rangle \geq g(F(\mathbf{w})).$$

- For $\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R)^c$, as $\chi_F, \chi_\Phi \equiv 1$ identically in this set, we have by Lemma 38

$$\left\langle \nabla\tilde{\Phi}(\mathbf{w}), \nabla\tilde{F}(\mathbf{w})\right\rangle = \left\langle \nabla\Phi_2(\mathbf{w}), \nabla\tilde{F}(\mathbf{w})\right\rangle \geq g(F(\mathbf{w})).$$

- For $\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R-1)^c \cap \mathbb{B}(\mathbf{w}^\star, R)$, we have

$$\nabla\tilde{\Phi}(\mathbf{w}) = \chi_\Phi(\mathbf{w})\nabla\Phi_2(\mathbf{w}) + (1-\chi_\Phi(\mathbf{w}))\nabla\Phi(\mathbf{w}) + \nabla\chi_\Phi(\mathbf{w})\Phi_2(\mathbf{w}) - \nabla\chi_\Phi(\mathbf{w})\Phi(\mathbf{w}).$$

  First let $L_{b,1}$ denote the minimum value of F in $\mathbb{B}(\mathbf{w}^\star, R-1)^c \cap \mathbb{B}(\mathbf{w}^\star, R)$. Note $L_{b,1} \geq l_b$ by assumption that $\mathbb{B}(\mathcal{W}^\star, r(l_b)) \subseteq \mathbb{B}(\mathbf{w}^\star, R-1)$. This means

$$\begin{aligned}\left\langle \nabla\tilde{\Phi}(\mathbf{w}), \nabla\tilde{F}(\mathbf{w})\right\rangle = (1-\chi_\Phi(\mathbf{w}))\left\langle \nabla\Phi(\mathbf{w}), \nabla\tilde{F}(\mathbf{w})\right\rangle + \chi_\Phi(\mathbf{w})\left\langle \nabla\Phi_2(\mathbf{w}), \nabla\tilde{F}(\mathbf{w})\right\rangle \\ + (\Phi_2(\mathbf{w}) - \Phi(\mathbf{w}))\left\langle \nabla\chi_\Phi(\mathbf{w}), \nabla\tilde{F}(\mathbf{w})\right\rangle. \qquad (44)\end{aligned}$$

  Note in this region,

$$\nabla\tilde{F}(\mathbf{w}) = \nabla F(\mathbf{w}) + \nabla\chi_F(\mathbf{w}) \cdot \lambda_{\text{REG}}(\|\mathbf{w}-\mathbf{w}^\star\|^2 + 1) + \chi_F(\mathbf{w}) \cdot 2\lambda_{\text{REG}}\|\mathbf{w}-\mathbf{w}^\star\|(\mathbf{w}-\mathbf{w}^\star).$$

  Also recall that

$$\nabla\chi_\Phi(\mathbf{w}) = p'_\Phi(\|\mathbf{w}-\mathbf{w}^\star\| - (R-1))\frac{\mathbf{w}-\mathbf{w}^\star}{\|\mathbf{w}-\mathbf{w}^\star\|}, \nabla\chi_F(\mathbf{w}) = p'_F(\|\mathbf{w}-\mathbf{w}^\star\| - (R-1))\frac{\mathbf{w}-\mathbf{w}^\star}{\|\mathbf{w}-\mathbf{w}^\star\|}.$$

  Define

$$\begin{aligned}
A &= \langle \nabla\Phi(\mathbf{w}), \nabla F(\mathbf{w})\rangle \geq g(F(\mathbf{w})) \geq 0,\\
B_1 &= \frac{p'_F(\|\mathbf{w}-\mathbf{w}^\star\| - (R-1))}{\|\mathbf{w}-\mathbf{w}^\star\|}\lambda_{\text{REG}}(\|\mathbf{w}-\mathbf{w}^\star\|^2+1)\langle \nabla\Phi(\mathbf{w}), \mathbf{w}-\mathbf{w}^\star\rangle,\\
B_2 &= \chi_F(\mathbf{w}) \cdot 2\lambda_{\text{REG}}\|\mathbf{w}-\mathbf{w}^\star\|\langle \nabla\Phi(\mathbf{w}), \mathbf{w}-\mathbf{w}^\star\rangle,\\
C_1 &= c_{\text{WGT}}\lambda_{\text{REG}}(\|\mathbf{w}-\mathbf{w}^\star\|^2+1)\langle \nabla\chi_F(\mathbf{w}), \mathbf{w}-\mathbf{w}^\star\rangle\\
&= c_{\text{WGT}}\lambda_{\text{REG}}(\|\mathbf{w}-\mathbf{w}^\star\|^2+1)\|\mathbf{w}-\mathbf{w}^\star\|p'_F(\|\mathbf{w}-\mathbf{w}^\star\| - (R-1)) \geq 0,\\
&= B_1 c_{\text{WGT}}\frac{\|\mathbf{w}-\mathbf{w}^\star\|^2}{\langle \nabla\Phi(\mathbf{w}), \mathbf{w}-\mathbf{w}^\star\rangle}, \qquad (45)\\
C_2 &= 2c_{\text{WGT}}\lambda_{\text{REG}}\chi_F(\mathbf{w})\|\mathbf{w}-\mathbf{w}^\star\|^3 \geq 0\\
&= B_2 c_{\text{WGT}}\frac{\|\mathbf{w}-\mathbf{w}^\star\|^2}{\langle \nabla\Phi(\mathbf{w}), \mathbf{w}-\mathbf{w}^\star\rangle}, \qquad (46)\\
C_3 &= c_{\text{WGT}}\langle \nabla F(\mathbf{w}), \mathbf{w}-\mathbf{w}^\star\rangle \geq 0.
\end{aligned}$$

  It is clear that $C_1, C_2 \geq 0$, and $C_3 \geq 0$ follows by Assumption A.2. In the above, A, C_1, C_2 are favorable terms, and B_1, B_2 are terms that could be negative that we must control.

Recalling the definition of Φ_2, that for $\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R)$ we have $\Phi(\mathbf{w}) \le M$, and furthermore using Corollary 39, we obtain from (44) that

$$\begin{aligned}
\langle \nabla \tilde{\Phi}(\mathbf{w}), \nabla \tilde{F}(\mathbf{w}) \rangle &\ge (1 - \chi_\Phi(\mathbf{w}))(A + B_1 + B_2) \\
&\quad + \left(c_{\text{WGT}} \chi_\Phi(\mathbf{w}) + \frac{M p'_\Phi(\|\mathbf{w} - \mathbf{w}^\star\| - (R-1))}{R} \right) \langle \nabla \tilde{F}(\mathbf{w}), \mathbf{w} - \mathbf{w}^\star \rangle \\
&\ge (1 - \chi_\Phi(\mathbf{w}))(A + B_1 + B_2) + c_{\text{WGT}} \chi_\Phi(\mathbf{w}) \langle \nabla \tilde{F}(\mathbf{w}), \mathbf{w} - \mathbf{w}^\star \rangle \\
&= (1 - \chi_\Phi(\mathbf{w}))(A + B_1 + B_2) + \chi_\Phi(\mathbf{w})(C_1 + C_2 + C_3).
\end{aligned}$$

We aim to find a lower bound on the above. We break into cases:

1. Suppose $\|\mathbf{w} - \mathbf{w}^\star\| < R - 1 + t_{\text{THRES},F}$. In this case by Corollary 39, it remains to lower bound $(1 - \chi_\Phi(\mathbf{w}))(A + B_1 + B_2)$ by a positive constant. (This is where it becomes very useful to have independent interpolators χ_F, χ_Φ.)

   By construction of χ_Φ, for $\|\mathbf{w} - \mathbf{w}^\star\| < R - 1 + t_{\text{THRES},F}$, recall we have

   $$\chi_\Phi(\mathbf{w}) = p_\Phi(\mathbf{w} - \mathbf{w}^\star - (R-1)) \le \frac{1}{2}.$$

   That is, we still 'weight' Φ substantially in the construction of $\tilde{\Phi}$. Furthermore for such $\mathbf{w}$, recall by (42) that

   $$\begin{aligned}
   F(\mathbf{w}) &\ge g(L_{b,1}). \\
   p'_F(\|\mathbf{w} - \mathbf{w}^\star\| - (R-1)) &\le c_F. \\
   \chi_F(\|\mathbf{w} - \mathbf{w}^\star\| - (R-1)) &= p_F(\|\mathbf{w} - \mathbf{w}^\star\| - (R-1)) \le c_F.
   \end{aligned}$$

   That is, we do not weight the regularizer much yet.

   Thus we obtain

   $$\begin{aligned}
   |B_1| &\le p'_F(\|\mathbf{w} - \mathbf{w}^\star\| - (R-1)) \cdot \lambda_{\text{REG}}(R^2 + 1)\rho_\Phi(M) \le \frac{1}{4} g(L_{b,1}). \\
   |B_2| &\le p_F(\|\mathbf{w} - \mathbf{w}^\star\| - (R-1)) \cdot 2\lambda_{\text{REG}} R^2 \rho_\Phi(M) \le \frac{1}{4} g(L_{b,1}).
   \end{aligned}$$

   Consequently we have

   $$A + B_1 + B_2 \ge g(F(\mathbf{w})) - \frac{1}{2} g(L_{b,1}) \ge \frac{1}{2} g(F(\mathbf{w})) + \frac{1}{2} g(L_{b,1}) - \frac{1}{2} g(L_{b,1}) = \frac{1}{2} g(F(\mathbf{w})).$$

   Recalling $C_1, C_2, C_3 \ge 0$, we obtain

   $$\langle \nabla \tilde{\Phi}(\mathbf{w}), \nabla \tilde{F}(\mathbf{w}) \rangle \ge (1 - \chi_\Phi(\mathbf{w}))(A + B_1 + B_2) \ge \frac{1}{4} g(F(\mathbf{w})) \ge \frac{1}{4} g(L_{b,1}).$$

2. Suppose $\|\mathbf{w} - \mathbf{w}^\star\| \ge R - 1 + t_{\text{THRES},F}$. In this case $A + B_1 + B_2 < 0$ is possible. The benefit however is that c_{WGT} comes into play, and allows for C_1, C_2 to dominate. The relations (45), (46) between B_1, C_1 and B_2, C_2 earlier, that $\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R)$ in this case, and the choice of c_{WGT} together imply that

   $$B_2 + C_2 \ge 0, B_1 + \frac{C_1}{2} \ge 0.$$

Notice here by construction of χ_Φ that in this case, we have $\chi_\Phi(\mathbf{w}) \geq \frac{1}{2}$. Consequently, we have

$$\begin{aligned}\left\langle \nabla\tilde{\Phi}(\mathbf{w}), \nabla\tilde{F}(\mathbf{w})\right\rangle &\geq (1-\chi_\Phi(\mathbf{w}))(A+B_1+B_2)+\chi_\Phi(\mathbf{w})(C_1+C_2+C_3)\\ &\geq \frac{1}{2}(B_1+B_2)+\frac{1}{2}(C_1+C_2)\\ &\geq \frac{1}{4}C_1+\frac{1}{4}C_2+\frac{1}{2}B_1 \geq \frac{1}{4}C_1.\end{aligned}$$

By choice of c_{WGT}, and since

$$p_F'(\|\mathbf{w}-\mathbf{w}^\star\|-(R-1)) \geq c_F \text{ for } \|\mathbf{w}-\mathbf{w}^\star\| \geq R-1+t_{\mathrm{THRES},F},$$

we have for such $\mathbf{w}$,

$$\frac{1}{4}C_1 \geq \frac{1}{4}c_{\mathrm{WGT}}\lambda_{\mathrm{REG}}((R-1)^2+1)(R-1)c_F \geq \frac{1}{4}g(L_{b,1}).$$

This last step follows by definition of c_{WGT}.

Thus in either case, we have

$$\left\langle \nabla\tilde{\Phi}(\mathbf{w}), \nabla\tilde{F}(\mathbf{w})\right\rangle \geq \frac{1}{4}g(L_{b,1}).$$

Putting these cases together, we obtain:

1. For $F(\mathbf{w}) \geq M$, then we must have $\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R)^c$ and so
$$\left\langle \nabla\tilde{\Phi}(\mathbf{w}), \nabla\tilde{F}(\mathbf{w})\right\rangle \geq g(F(\mathbf{w})).$$
2. For $F(\mathbf{w}) \in [l_b, M)$, then as F is non-decreasing and as $L_{b,1} \geq l_b$,
$$\left\langle \nabla\tilde{\Phi}(\mathbf{w}), \nabla\tilde{F}(\mathbf{w})\right\rangle \geq \frac{1}{4}g(l_b).$$
3. For $F(\mathbf{w}) \leq l_b$, we must have $\mathbf{w} \in \mathbb{B}(\mathbf{w}^\star, R-1)$ and so
$$\left\langle \nabla\tilde{\Phi}(\mathbf{w}), \nabla\tilde{F}(\mathbf{w})\right\rangle \geq g(F(\mathbf{w})).$$

We now construct a non-decreasing, infinitely differentiable function $\tilde{h}$ analogously to the definition of $\tilde{g}$ from Subsection D.1. Notice $\frac{1}{4}g(L_{b,1}) \leq g(M)$ as $L_{b,1} \leq M$ and as g is non-decreasing. Now for some small constant $1 > \delta > 0$, we can interpolate to create $\tilde{h}$ as follows:

$$\tilde{h}(x) = \begin{cases} \frac{1}{8}g(x) = \frac{1}{8}\lambda x & : x \leq l_b \\ \text{smooth interpolation to } \frac{1}{4}g(l_b) & : l_b < x < l_b+\delta \\ \frac{1}{4}g(l_b) & : l_b+\delta \leq x \leq M \\ \text{smooth interpolation to } \lambda x & : M < x < M+\delta \\ \lambda x & : M+\delta \leq x \end{cases}. \tag{47}$$

These interpolators can be defined analogously as in the definition of $\tilde{g}$, from Subsection D.1, so that $\tilde{h}$ is non-decreasing and differentiable, and so that $\tilde{h}(x) \leq \lambda x = g(x)$ for $x \in [M, M+\delta]$ (because we took M

so that we have $\lambda x \geq \frac{1}{4}g(L_{b,1}) + 1 \geq \frac{1}{4}g(l_b) + 1$ for $x \geq M$), and $\tilde{h}(x) \leq \frac{1}{8}g(l_b) \leq \frac{1}{4}g(l_b)$ for $x \in [\tilde{l}, \tilde{l} + \delta]$. Moreover, note $\tilde{h}(x) = \lambda x$ for $x \geq \tilde{M} + 1$.

Noting $\tilde{h}(x) \geq 0$, define

$$m'_{\text{NEW}} = \lambda, b'_{\text{NEW}} = \lambda(M+1)', \tag{48}$$

where M is defined as per (40). Consequently we always have $\tilde{h}(x) \geq m'_{\text{NEW}}x - b'_{\text{NEW}}$.

Therefore, for all $\mathbf{w} \in \mathbb{R}^d$ we have

$$\left\langle \nabla\tilde{\Phi}(\mathbf{w}), \nabla\tilde{F}(\mathbf{w})\right\rangle \geq \tilde{h}(F(\mathbf{w})). \tag{49}$$

We can also check now similarly to Part 1 of Subsection D.1 that

$$\left\|\nabla^2\tilde{\Phi}\right\|_{\text{op}} \leq L' + B_\Phi\left(R^2 c_{\text{WGT}} + 4M\right) + \left(c_{\text{WGT}} + L'\right)\cdot 1 + 2B_\Phi\left(L' + Rc_{\text{WGT}}\right),$$

where

$$L' = \sup_{\mathbf{w}\in\mathbb{B}(\mathbf{w}^\star, R)} \rho_\Phi(\Phi(\mathbf{w})). \tag{50}$$

Consequently, $\tilde{\Phi}$ is again $\tilde{L}$-smooth, where we now define

$$\tilde{L} := \left(L' + B_\Phi\left(R^2 c_{\text{WGT}} + 4M\right) + \left(c_{\text{WGT}} + L'\right)\cdot 1 + 2B_\Phi\left(L' + Rc_{\text{WGT}}\right)\right) \vee 2b'_{\text{NEW}} \vee 1. \tag{51}$$

Part 2: Proving a PI with the same idea as before. From here, the finish is analogous to the proof of Theorem 12. We omit straightforward details that are checked verbatim as there. Because of (49), letting $h(x) = \tilde{h}(x) + B$ where $B = \tilde{L}$, Lemma 27 gives for any test function f:

$$\int f(\mathbf{w})^2 \frac{\tilde{h}(F(\mathbf{w}))}{\tilde{h}(F(\mathbf{w})) + \tilde{L}} \mathrm{d}\mu_\beta \leq \frac{1}{\beta}\int\left(\|\nabla f(\mathbf{w})\|^2 + \frac{f(\mathbf{w})^2}{h(\mathbf{w})^2}\left\|\nabla\tilde{\Phi}(\mathbf{w})\right\|^2 - \frac{f(\mathbf{w})^2}{h(\mathbf{w})^2}\left\langle\nabla h(\mathbf{w}), \nabla\tilde{\Phi}(\mathbf{w})\right\rangle\right)\mathrm{d}\mu_\beta + \frac{1}{\beta}\int f(\mathbf{w})^2 \frac{\left|\Delta\tilde{\Phi}(\mathbf{w})\right|}{h(\mathbf{w})}\mathrm{d}\mu_\beta. \tag{52}$$

Step a: Upper bounding intermediate terms. Using $\tilde{L}$-smoothness of $\tilde{\Phi}$, and that $\tilde{h}(x) \geq m'_{\text{NEW}}x - b'_{\text{NEW}}$, $\tilde{L}/2 \geq b'_{\text{NEW}}$, $F(\mathbf{w}) \geq r_2\|\mathbf{w} - \mathbf{w}^\star\|$, we obtain analogously to Step a in Subsection D.1 that

$$\frac{\left\|\nabla\tilde{\Phi}(\mathbf{w})\right\|^2 - \left\langle\nabla h(\mathbf{w}), \nabla\tilde{\Phi}(\mathbf{w})\right\rangle}{h(\mathbf{w})^2} \leq C',$$

where now

$$C' := \frac{8\left(R^2 + 8Mr_1\right)}{r_1\tilde{L}} \vee \frac{2\tilde{L}}{r_1 r_2^2 m'^2_{\text{NEW}}} \vee \frac{64M}{\tilde{L}}., \tag{53}$$

where $\tilde{L}$ is defined in (51), $m'_{\text{NEW}} = \lambda$, and M is defined in (40).

Step b: Finishing the proof of PI identically to before. Consider an arbitrary test function ψ and define f in terms of ψ identically as in Subsection D.1, (20).

Now using C' to upper bound the right hand side of (52), we obtain

$$\int f(\mathbf{w})^2 \frac{\tilde{h}(F(\mathbf{w}))}{\tilde{h}(F(\mathbf{w})) + \tilde{L}} \mathrm{d}\mu_\beta \leq \frac{1}{\beta}\int\|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\int f(\mathbf{w})^2\left(d + C'\right)\mathrm{d}\mu_\beta.$$

The only difference is the $\tilde{h}$ rather than $\tilde{g}$ in the left hand side, and that now C' is defined in (53), rather than (22).

Now recalling that $\tilde{h}$ is non-decreasing, we obtain from the above that

$$\int f(\mathbf{w})^2 \frac{\tilde{h}(l_b)}{\tilde{h}(l_b)+\tilde{L}} \mathrm{d}\mu_\beta \le \frac{1}{\beta}\int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\int f(\mathbf{w})^2\big(d+C'\big)\mathrm{d}\mu_\beta + \frac{\tilde{h}(l_b)}{\tilde{h}(l_b)+\tilde{L}}\int_{\mathcal{U}} f(\mathbf{w})^2 \mathrm{d}\mu_\beta.$$

An analogous manipulation using Assumption 3.1 to upper bound $\int_{\mathcal{U}} f(\mathbf{w})^2\mathrm{d}\mu_\beta$, using the choice of α, now proves

$$\begin{aligned}\int f(\mathbf{w})^2 \frac{\tilde{h}(l_b)}{\tilde{h}(l_b)+\tilde{L}} \mathrm{d}\mu_\beta &\le \frac{1}{\beta}\int \|\nabla f(\mathbf{w})\|^2 \mathrm{d}\mu_\beta + \frac{1}{\beta}\int f(\mathbf{w})^2\big(d+C'\big)\mathrm{d}\mu_\beta \\ &\quad + \frac{\tilde{h}(l_b)}{\tilde{h}(l_b)+\tilde{L}}\cdot \mathsf{C}_{\text{PI, LOCAL}}\int \|\nabla f(\mathbf{w})\|^2\mathrm{d}\mu_\beta.\end{aligned}$$

If $\beta \ge 2\left(1+\frac{\tilde{L}}{\tilde{h}(l_b)}\right)(d+C') = \Omega(d)$, where $\tilde{h}$ is as per (47), C' is as per (53), $\tilde{L}$ is as per (51), we obtain

$$\mathbb{V}_{\mu_\beta}[\psi] \le \left(2\mathsf{C}_{\text{PI, LOCAL}} + \frac{2}{\beta}\left(1+\frac{\tilde{L}}{\tilde{h}(l_b)}\right)\right)\int \|\nabla\psi\|^2\mathrm{d}\mu_\beta.$$

Recalling ψ is an arbitrary test function, this gives the desired Poincaré Inequality. We furthermore verified that $\hat{F}$ is $O(1)$-smooth in Lemma 37, so this finishes the proof. ■

Remark 40 *Note if we instead have an upper bound of the form $\|\nabla F(\mathbf{w})\| \le L(\|\mathbf{w}-\mathbf{w}^\star\|^s+1)$ rather than smoothness, one can instead add regularization in the form $\lambda_{\text{REG}}(\|\mathbf{w}-\mathbf{w}^\star\|^{s+1}+1)$. To capture other forms of $g(\cdot)$, one can perform similar ideas of lower bounding $g(x)$ by a $\tilde{g}(x)$ that grows linearly for large enough x, as done in Subsection D.1. One can also tighten the PI to an LSI as in Subsection D.1. These details follow the exact same argument as in Subsection D.1 and are straightforward to verify.*

Remark 41 *Notice to construct ∇F, all the problem-dependent parameters used in the construction can be computed with oracle access to F, knowledge of $\mathbf{w}^\star$, R, except for $\rho_\Phi(M)$ (to define c_F) and L (to define λ_{REG}). However, for $\rho_\Phi(M)$, L, it suffices to use a upper bound on them, as can be seen through the above proof. Consequently we can construct a suitable $\hat{F}$ via appropriate cross-validation on these parameters.*

Appendix E. Technical Helper Results

Lemma 42 (Lemma 2.1, Srebro et al. (2010)) *If some G is non-negative and L-smooth, then*

$$\|\nabla G(\mathbf{w})\| \le \sqrt{4LG(\mathbf{w})}.$$

Lemma 43 *Suppose F is L-Hölder continuous with parameter $s \in (0,1]$. Let $M = \int \|\cdot\|\mathrm{d}\mu_\beta$. Additionally define $\hat{F}(\mathbf{w}) = F(\mathbf{w}) + \frac{\gamma}{2\beta}\max(0,\|\mathbf{w}\|-R)^2$ for $\gamma > 0$, $\hat{\mu}_\beta = e^{-\beta\hat{F}}/Z$. With initialization $\pi_0 \sim \mathcal{N}\left(\vec{\mathbf{0}}, \frac{1}{2\beta L+\gamma}I_d\right)$, we have the following:*

$$\ln\big(\chi^2(\pi_0\|\mu_\beta)+1\big), \mathsf{KL}(\pi_0\|\mu_\beta) \le \beta L + \beta F(\vec{\mathbf{0}}) + 2 + \frac{d}{2}\ln\big(4M^2(\beta L+\gamma/2)\big),$$

$$\ln\big(\chi^2(\pi_0\|\hat{\mu}_\beta)+1\big), \mathsf{KL}(\pi_0\|\hat{\mu}_\beta) \le \beta L + \beta F(\vec{\mathbf{0}}) + 2 + \frac{d}{2}\ln\big(4\hat{M}^2(\beta L+\gamma/2)\big).$$

Proof. Since Rényi divergence (for the definition, see e.g. Chewi (2024)) is increasing in its order, and as KL divergence is Rényi divergence of order 1 and $\ln(\chi^2 + 1)$ is Rényi divergence of order 2, it suffices to show these upper bounds for the Rényi divergence of order ∞, $\mathcal{R}_\infty(\cdot\|\cdot)$. This is the supremum of the log ratio of the probability density functions. Now the proof follows by exactly the same argument as the proof of Lemmas 31 and 32 from Chewi et al. (2024). We highlight it here by proving the second upper bound. Let $V = \beta F$, $\hat{V} = \beta \hat{F}$. Then we can compare the ratio of their unnormalized densities:

$$\begin{aligned}\exp\Big(\hat{V}(\mathbf{w}) - \Big(L\beta + \frac{\gamma}{2}\Big)\|\mathbf{w}\|^2\Big) &\le \exp\Big(\hat{V}(\mathbf{w}) - \hat{V}(\vec{\mathbf{0}}) + \hat{V}(\vec{\mathbf{0}}) - \Big(\beta L + \frac{\gamma}{2}\Big)\|\mathbf{w}\|^2\Big)\\ &\le \exp\Big(\beta L\|\mathbf{w}\|^{s+1} + \frac{\gamma}{2}\max\{0, \|\mathbf{w}\| - R\}^2 + \beta F(\vec{\mathbf{0}}) - \Big(L\beta + \frac{\gamma}{2}\Big)\|\mathbf{w}\|^2\Big)\\ &\le \exp\big(\beta L + \beta F(\vec{\mathbf{0}})\big).\end{aligned}$$

Here we used the inequality $x^{s+1} \le x^2 + 1$ for all $x \ge 0$ (as $s \le 1$) and $\hat{V}(\mathbf{w}) - \hat{V}(\vec{\mathbf{0}}) = \beta\big(F(\mathbf{w}) - F(\vec{\mathbf{0}})\big) + \frac{\gamma}{2}\max\{0, \|\mathbf{w}\| - R\}^2 \le \beta L\|\mathbf{w}\|^{s+1} + \frac{\gamma}{2}\max\{0, \|\mathbf{w}\| - R\}^2$.

Now analogously to the proof of Lemma 31 of Chewi et al. (2024), we compare the partition functions, arguing through the intermediate quantity $\int \exp\big(-\hat{V}(\mathbf{w}) - \delta\|\mathbf{w}\|^2\big)\mathrm{d}\mathbf{w}$:

$$\frac{\int \exp\big(-\hat{V}(\mathbf{w}) - \delta\|\mathbf{w}\|^2\big)\mathrm{d}\mathbf{w}}{\int \exp\big(-\hat{V}(\mathbf{w})\big)\mathrm{d}\mathbf{w}} \ge \frac{1}{2}\exp\big(-4\delta\hat{M}^2\big), \frac{\int \exp\big(-\hat{V}(\mathbf{w}) - \delta\|\mathbf{w}\|^2\big)\mathrm{d}\mathbf{w}}{\left(\frac{\pi}{\beta L + \gamma/2}\right)^{d/2}} \le \left(\frac{\beta L + \gamma/2}{\delta}\right)^{d/2}.$$

Taking $\delta = \frac{1}{4\hat{M}^2}$ and rearranging the above gives

$$\mathcal{R}_\infty\big(\pi_0\|\hat{\mu}_\beta\big) \le \beta L + \beta F(\vec{\mathbf{0}}) + 2 + \frac{d}{2}\ln\big(4\hat{M}^2(\beta L + \gamma/2)\big).$$

For the first upper bound, we do the same steps with V in place of $\hat{V}$. The first upper bound still holds, and the second two inequalities comparing the partition functions still hold, except $\hat{M}$ is replaced by M instead. Taking $\delta = \frac{1}{4M^2}$, we obtain the first inequality. ∎

Remark 44 *For an upper bound on M and $\hat{M}$, note if F is L-smooth and dissipative, that is $\langle \mathbf{w}, \nabla F(\mathbf{w})\rangle \ge m\|\mathbf{w}\|^2 - b$ for $m, b > 0$, then following the notation from Theorem 12, Cauchy-Schwartz gives that*

$$M^2 \le S \le \frac{b + d/\beta}{m} = O(1).$$

The bound on S follows from Raginsky et al. (2017). If F is dissipative with parameters m, b it is easy to check $\hat{F}$ is also dissipative with the same parameters, so we also have the same upper bound on $\hat{M}$. Notice also for $F = \|\mathbf{w}\|^\alpha$ and $\beta = \Omega(d)$ that $M = O(1)$. Therefore, we believe it is reasonable to suppose the right hand side of the above two lines is $\tilde{O}(\beta)$ for $\beta = \Omega(d)$.[21]

Remark 45 *As will be clear in the following proof, it is also possible to replace each instance of $\mathbf{w}$ with $\mathbf{w} - \mathbf{w}^\star$ for a fixed $\mathbf{w}^\star \in \mathcal{W}^\star$, if we know such a $\mathbf{w}^\star$. Our initialization then changes to Gaussian initialization centered at $\mathbf{w}^\star$. This can be done to give somewhat better bounds, but we do not pursue it for simplicity.*

Lemma 46 *Suppose F satisfies Assumption D.1. Taking $\pi_0(\mathbf{w}) \propto \exp\Big(-2\|\mathbf{w}\|^{2s_3'}\Big)$ where $s_3' = \max(s_3 + \frac{1}{2}, r + 1)$, we have*

$$\mathsf{KL}(\pi_0\|\mu_\beta), \mathsf{KL}(\pi_0\|\hat{\mu}_\beta) \le \tilde{O}(d\beta).$$

21. Since we are in the low temperature setting corresponding to optimization, the norm is a β factor *smaller* than in the standard sampling setting.

Here $\hat{\mu}_\beta$ comes from Theorem 3, Lytras and Mertikopoulos (2024); it is defined explicitly in our proof of Theorem 12.

Proof. First notice by Assumption D.1, we can check that for some $L_1, L_2 > 0$, we have $F(\mathbf{w}) \leq L_1\|\mathbf{w}\|^{2s_3+1} + L_2$. Thus $F(\mathbf{w}), F(\mathbf{w}) + \frac{\eta}{\beta}\|\mathbf{w}\|^{2r+2} \leq L_1\|\mathbf{w}\|^{2s_3'} + L_2$ where $s_3' = \max(s_3 + \frac{1}{2}, r+1)$. Now we adopt the proof of Lemma 5, Raginsky et al. (2017). Analogously to how C.11 was derived there, we have

$$\mathsf{KL}(\pi_0\|\mu_\beta) \leq \log\|\pi_0\|_\infty + \log\Lambda + \beta\int_{\mathbb{R}^d}\pi_0(\mathbf{w})F(\mathbf{w})\mathrm{d}\mathbf{w}, \tag{54}$$

where Λ denotes the partition function of μ_β. We upper bound each part of the above sum:

- The partition function: By the second part of Assumption 3.2, we have

$$\begin{aligned}\Lambda &= \int_{\mathbb{R}^d} e^{-\beta F(\mathbf{w})}\mathrm{d}\mathbf{w}\\ &\leq e^{\beta\sup_{\mathbf{w}\in\mathbb{B}(\mathbf{w}^\star,R)}F(\mathbf{w})}\int_{\mathbb{R}^d} e^{-\beta r_2\|\mathbf{w}-\mathbf{w}^\star\|}\mathrm{d}\mathbf{w}\\ &= e^{\beta\sup_{\mathbf{w}\in\mathbb{B}(\mathbf{w}^\star,R)}F(\mathbf{w})}\frac{2\pi^{d/2}}{\Gamma(d/2)}(\beta r_2)^{-d}\Gamma(d)\\ &\leq e^{\beta\sup_{\mathbf{w}\in\mathbb{B}(\mathbf{w}^\star,R)}F(\mathbf{w})}\cdot\frac{4\pi^{d/2}\cdot d^d\sqrt{\pi}}{(\beta r_2)^d}.\end{aligned}$$

  Here $\Gamma(\cdot)$ denotes the Gamma function. We evaluated the integral by Lemma 8.5 of Chen et al. (2024), and used straightforward properties of $\Gamma(\cdot)$ in the above.

- The ∞ norm: Since $\pi_0(\mathbf{w}) \propto \exp\left(-2\|\mathbf{w}\|^{2s_3'}\right)$, it follows that its normalizing constant is

$$Z = \int_{\mathbb{R}^d}\exp\left(-2\|\mathbf{w}\|^{2s_3'}\right)\mathrm{d}\mathbf{w} = \frac{2\pi^{d/2}}{\Gamma(d/2)}\cdot\frac{1}{2s_3'}2^{-\frac{d-2}{2s_3'}}\Gamma\left(\frac{d}{2s_3'}\right) \geq \frac{\pi^{d/2}}{2\pi^{1/2}s_3'd^{d/2}2^{\frac{d-2}{2s_3'}}}.$$

  The computation of this integral follows from analogous steps as in Lemmas 5.1 and 8.5, Chen et al. (2024) (there the result is stated for a particular range on s_3', but this is not needed). It follows that for all $\mathbf{w}\in\mathbb{R}^d$,

$$\log\pi_0 = -2\|\mathbf{w}\|^{2s} - \log Z \leq -\log Z \leq \log(2s_3'\pi^{1/2}) + \frac{d}{2}\log\left(\frac{d2^{2s_3'}}{\pi}\right).$$

- The last term: Since $F(\mathbf{w}) \leq L_1\|\mathbf{w}\|^{2s_3'} + L_2$,

$$\int_{\mathbb{R}^d}\pi_0(\mathbf{w})F(\mathbf{w})\mathrm{d}\mathbf{w} \leq \int_{\mathbb{R}^d}\pi_0(\mathbf{w})F(\mathbf{w})\mathrm{d}\mathbf{w} \leq L_1\int_{\mathbb{R}^d}\pi_0(\mathbf{w})\|\mathbf{w}\|^{2s_3'}\mathrm{d}\mathbf{w} + L_2.$$

  By Jensen's Inequality, we have

$$\int_{\mathbb{R}^d}\pi_0(\mathbf{w})\|\mathbf{w}\|^{2s_3'} = \mathbb{E}_{\pi_0}\left[\log\exp\left\{\|\mathbf{w}\|^{2s_3'}\right\}\right] \leq \log\mathbb{E}_{\pi_0}\left[\exp\left\{\|\mathbf{w}\|^{2s_3'}\right\}\right].$$

  Let Z denote the normalizing constant of π_0, as in the above. Note by choice of π_0,

$$\begin{aligned}\mathbb{E}_{\pi_0}\left[\exp\left(\|\mathbf{w}\|^{2s_3'}\right)\right] &= \frac{1}{Z}\int\exp\left(\|\mathbf{w}\|^{2s_3'} - 2\|\mathbf{w}\|^{2s_3'}\right)\mathrm{d}\mathbf{w}\\ &= \frac{1}{Z}\int\exp\left(-\|\mathbf{w}\|^{2s_3'}\right)\mathrm{d}\mathbf{w}\end{aligned}$$

$$= \frac{\frac{2\pi^{d/2}}{\Gamma(d/2)} \cdot \frac{1}{2s_3'}\Gamma\left(\frac{d}{2s_3'}\right)}{\frac{2\pi^{d/2}}{\Gamma(d/2)} \cdot \frac{1}{2s_3'}2^{-\frac{d-2}{2s_3'}}\Gamma\left(\frac{d}{2s_3'}\right)} = e^{\ln 2 \cdot \frac{d-2}{2s_3'}}.$$

Here, we evaluated the above integral analogously to how we computed Z. Putting all this together yields

$$\int_{\mathbb{R}^d} \pi_0(\mathbf{w})F(\mathbf{w})\mathrm{d}\mathbf{w} \le L_1 \cdot \frac{d-2}{s_3'} + L_2.$$

Putting all these steps together yields

$$\begin{aligned}
&\mathsf{KL}(\pi_0 \| \mu_\beta) \\
&\le \log\|\pi_0\|_\infty + \log\Lambda + \beta\int_{\mathbb{R}^d} \pi_0(\mathbf{w})F(\mathbf{w})\mathrm{d}\mathbf{w} \\
&\le \log(2s_3'\pi^{1/2}) + \frac{d}{2}\log\left(\frac{d2^{2s_3'}}{\pi}\right) + \beta \sup_{\mathbf{w}\in\mathbb{B}(\mathbf{w}^\star,R)} F(\mathbf{w}) + d\log\left(\frac{4\pi^{\frac{1}{2}+\frac{1}{2d}}d}{\beta r_2}\right) + \beta\left(L_1 \cdot \frac{d-2}{s_3'} + L_2\right) \\
&= \tilde{O}(d\beta).
\end{aligned}$$

The calculation for $\mathsf{KL}(\pi_0 \| \hat{\mu}_\beta)$ follows from an analogous argument, using (54). We just replace $F(\mathbf{w})$ by $F(\mathbf{w}) + \frac{\eta}{\beta}\|\mathbf{w}\|^{2r+2}$, and thanks to the definition of s_3', all the bounds above go through. ■

Lemma 47 *We can construct a $\chi(\mathbf{w}) \in [0,1]$ such that:*

- *$\chi \equiv 0$ on $B(\mathbf{w}^\star, R)$ and $\chi \equiv 1$ on $B(\mathbf{w}^\star, R+1)^c$.*
- *$\chi(\mathbf{w})$ is differentiable to all orders.*
- *$\|\nabla\chi(\mathbf{w})\|, \|\nabla^2\chi(\mathbf{w})\|_{\mathrm{op}} \le B$ for some universal constant $B > 0$.*
- *$\langle\nabla\chi(\mathbf{w}), \nabla F(\mathbf{w})\rangle \ge 0$.*

Proof. The construction is to let

$$\chi(\mathbf{w}) = \begin{cases} 0 : \|\mathbf{w} - \mathbf{w}^\star\| \le R \\ 1 : \|\mathbf{w} - \mathbf{w}^\star\| \ge R + 1 \\ \frac{e^{-\frac{1}{(\|\mathbf{w}-\mathbf{w}^\star\|-R)^2}}}{e^{-\frac{1}{(\|\mathbf{w}-\mathbf{w}^\star\|-R)^2}} + e^{-\frac{1}{1-(\|\mathbf{w}-\mathbf{w}^\star\|-R)^2}}} : R < \|\mathbf{w} - \mathbf{w}^\star\| < R + 1. \end{cases}$$

Clearly $\chi \in [0,1]$ and also the first property is satisfied. The second property is satisfied because $\tilde{\chi}(x) = \frac{e^{-\frac{1}{x^2}}}{e^{-\frac{1}{x^2}} + e^{-\frac{1}{1-x^2}}}$ is infinitely differentiable on $(0,1)$, and $\tilde{\chi}(0) = 0$, $\tilde{\chi}(1) = 1$. In particular, on $(0,1)$, $e^{-\frac{1}{x^2}}$ and $e^{-\frac{1}{1-x^2}}$ are both infinitely differentiable, which can be verified by a straightforward induction argument, and their sum is lower bounded by a constant $[0,1]$. Therefore, the quotient $\tilde{\chi}(x)$ is infinitely differentiable. Therefore, $\tilde{\chi}$ interpolates between 0 and 1 on $(0,1)$ in an infinitely differentiable way. Because $R > 0$, the composition of $\tilde{\chi}$ and $\|\mathbf{w} - \mathbf{w}^\star\| - R$ is infinitely differentiable, as both these maps are.

For the next two properties, we directly do the calculation. They are both obvious when $\|\mathbf{w} - \mathbf{w}^\star\| \le R$ or $\|\mathbf{w} - \mathbf{w}^\star\| \ge R + 1$, so we check these two properties when $R < \|\mathbf{w} - \mathbf{w}^\star\| < R + 1$. We first prove the last property. We do so using the intuitive geometric approach of comparing the angle that $\nabla\chi(\mathbf{w})$ and $\nabla F(\mathbf{w})$ make with $\mathbf{w} - \mathbf{w}^\star$, and showing the sum of their angles is at most $\frac{\pi}{2}$.

First, by Assumption 3.2, we have for $R+1 > \|\mathbf{w}-\mathbf{w}^\star\| > R$ that

$$\frac{\langle \mathbf{w}-\mathbf{w}^\star, \nabla F(\mathbf{w})\rangle}{\|\mathbf{w}-\mathbf{w}^\star\|\|\nabla F(\mathbf{w})\|} \geq \frac{r_1 F(\mathbf{w})}{\|\mathbf{w}-\mathbf{w}^\star\|\|\nabla F(\mathbf{w})\|} \geq 0.$$

This means

$$\theta\langle \nabla F(\mathbf{w}), \mathbf{w}-\mathbf{w}^\star\rangle \leq \cos^{-1}(0) = \frac{\pi}{2}. \tag{55}$$

Notice $\nabla(\|\mathbf{w}-\mathbf{w}^\star\|) = \frac{\mathbf{w}-\mathbf{w}^\star}{\|\mathbf{w}-\mathbf{w}^\star\|}$. Thus, by Chain Rule, cwe have

$$\begin{aligned}
&\nabla\chi(\mathbf{w})\\
&= \frac{e^{-\frac{1}{(\|\mathbf{w}-\mathbf{w}^\star\|-R)^2}} \cdot \frac{2}{(\|\mathbf{w}-\mathbf{w}^\star\|-R)^3} \cdot \frac{\mathbf{w}-\mathbf{w}^\star}{\|\mathbf{w}-\mathbf{w}^\star\|}}{e^{-\frac{1}{(\|\mathbf{w}-\mathbf{w}^\star\|-R)^2}} + e^{-\frac{1}{1-(\|\mathbf{w}-\mathbf{w}^\star\|-R)^2}}}\\
&+ \frac{e^{-\frac{1}{(\|\mathbf{w}-\mathbf{w}^\star\|-R)^2}}\left(e^{-\frac{1}{(\|\mathbf{w}-\mathbf{w}^\star\|-R)^2}} \cdot \frac{2}{(\|\mathbf{w}-\mathbf{w}^\star\|-R)^3} \cdot \frac{\mathbf{w}-\mathbf{w}^\star}{\|\mathbf{w}-\mathbf{w}^\star\|} + e^{-\frac{1}{1-(\|\mathbf{w}-\mathbf{w}^\star\|-R)^2}} \cdot \frac{-2(\|\mathbf{w}-\mathbf{w}^\star\|-R)}{(1-(\|\mathbf{w}-\mathbf{w}^\star\|-R)^2)^2} \cdot \frac{\mathbf{w}-\mathbf{w}^\star}{\|\mathbf{w}-\mathbf{w}^\star\|}\right)}{\left(e^{-\frac{1}{(\|\mathbf{w}-\mathbf{w}^\star\|-R)^2}} + e^{-\frac{1}{1-(\|\mathbf{w}-\mathbf{w}^\star\|-R)^2}}\right)^2}.
\end{aligned}$$

Thus,

$$\nabla\chi(\mathbf{w}) = \tilde{p}(\|\mathbf{w}-\mathbf{w}^\star\| - R) \cdot \frac{\mathbf{w}-\mathbf{w}^\star}{\|\mathbf{w}-\mathbf{w}^\star\|},$$

where

$$\tilde{p}(x) = \frac{e^{-\frac{1}{x^2}} \cdot \frac{2}{x^3}}{e^{-\frac{1}{x^2}} + e^{-\frac{1}{1-x^2}}} + \frac{e^{-\frac{1}{x^2}}\left(e^{-\frac{1}{x^2}} \cdot \frac{2}{x^3} + e^{-\frac{1}{1-x^2}} \cdot \frac{-2x}{(1-x^2)^2}\right)}{(e^{-\frac{1}{x^2}} + e^{-\frac{1}{1-x^2}})^2}$$

is just a scalar. In Lemma 48, we prove $\tilde{p}(x) \geq 0$ for all $x \in [0,1]$, therefore

$$\langle \nabla\chi(\mathbf{w}), \mathbf{w}-\mathbf{w}^\star\rangle = \frac{\tilde{p}(\|\mathbf{w}-\mathbf{w}^\star\| - R)}{\|\mathbf{w}-\mathbf{w}^\star\|}\|\mathbf{w}-\mathbf{w}^\star\|^2 = \|\nabla\chi(\mathbf{w})\|\|\mathbf{w}-\mathbf{w}^\star\|.$$

Thus, the vectors $\nabla\chi(\mathbf{w}), \mathbf{w}-\mathbf{w}^\star$ are collinear and point in the same direction:

$$\theta\langle \nabla\chi(\mathbf{w}), \mathbf{w}-\mathbf{w}^\star\rangle = 0. \tag{56}$$

Combining (56) and (55), it is clear that $\theta\langle \nabla\chi(\mathbf{w}), \nabla F(\mathbf{w})\rangle \leq \frac{\pi}{2}$, hence $\langle \nabla\chi(\mathbf{w}), \nabla F(\mathbf{w})\rangle \geq 0$.

For the third property, we clearly only need to check it when $\|\mathbf{w}-\mathbf{w}^\star\| \in [R, R+1]$. The above calculation verifies it directly for the gradient Euclidean norm, as it shows that

$$\|\nabla\chi(\mathbf{w})\| = \tilde{p}(\|\mathbf{w}-\mathbf{w}^\star\| - R) \leq \sup_{t\in(0,1)} \tilde{p}(t).$$

We conclude this part for the gradient, noting $\tilde{p}$ is a univariate function with no explicit d dependence, which can be extended to be bounded and differentiable to all orders on $[0,1]$ (because $\lim_{t\to 0} e^{-1/t}\frac{1}{t^p} = 0$ for all $p < \infty$, and similarly for the limits to 1). For the Hessian operator norm, applying Chain Rule to the above shows

$$\begin{aligned}
\nabla^2\chi(\mathbf{w}) = \tilde{p}'(\|\mathbf{w}-\mathbf{w}^\star\| - R) \cdot \frac{1}{\|\mathbf{w}-\mathbf{w}^\star\|^2}(\mathbf{w}-\mathbf{w}^\star)(\mathbf{w}-\mathbf{w}^\star)^T\\
+ \tilde{p}(\|\mathbf{w}-\mathbf{w}^\star\| - R) \cdot \frac{1}{\|\mathbf{w}-\mathbf{w}^\star\|} I_d
\end{aligned}$$

$$- \tilde{p}(\|\mathbf{w}-\mathbf{w}^\star\|-R)\cdot\frac{1}{\|\mathbf{w}-\mathbf{w}^\star\|^2}\cdot\frac{1}{\|\mathbf{w}-\mathbf{w}^\star\|}(\mathbf{w}-\mathbf{w}^\star)(\mathbf{w}-\mathbf{w}^\star)^T.$$

The same rationale as before justifies that $\tilde{p}'(\cdot)$ is a univariate function with no explicit d dependence, which can be extended to be bounded and differentiable to all orders on $[0,1]$. Recalling $\|\mathbf{w}-\mathbf{w}^\star\|\in(R,R+1)$, it follows that $\tilde{p}'(\|\mathbf{w}-\mathbf{w}^\star\|-R)$ is upper bounded by universal constant $\sup_{t\in(0,1)}\tilde{p}'(t)<\infty$. Using the fact that

$$\left\|(\mathbf{w}-\mathbf{w}^\star)(\mathbf{w}-\mathbf{w}^\star)^T\right\|_{\text{op}}\le\|\mathbf{w}-\mathbf{w}^\star\|^2$$

when $R+1>\|\mathbf{w}-\mathbf{w}^\star\|>R$, we obtain

$$\begin{aligned}\left\|\nabla^2\chi(\mathbf{w})\right\|_{\text{op}} &\le \sup_{t\in(0,1)}\tilde{p}'(t)\cdot\frac{\|\mathbf{w}-\mathbf{w}^\star\|^2}{\|\mathbf{w}-\mathbf{w}^\star\|^2}+\sup_{t\in(0,1)}\tilde{p}(t)\cdot\frac{1}{R}+\sup_{t\in(0,1)}\tilde{p}(t)\cdot\frac{\|\mathbf{w}-\mathbf{w}^\star\|^2}{\|\mathbf{w}-\mathbf{w}^\star\|^3}\\ &\le \sup_{t\in(0,1)}\tilde{p}'(t)+2\sup_{t\in(0,1)}\tilde{p}(t).\end{aligned}$$

The last step follows as we have $R\ge1$ without loss of generality. The proof is complete. ∎

Lemma 48 *For $x\in[0,1]$, we have*

$$\tilde{p}(x)=\frac{e^{-\frac{1}{x^2}}\cdot\frac{2}{x^3}}{e^{-\frac{1}{x^2}}+e^{-\frac{1}{1-x^2}}}+\frac{e^{-\frac{1}{x^2}}\left(e^{-\frac{1}{x^2}}\cdot\frac{2}{x^3}+e^{-\frac{1}{1-x^2}}\cdot\frac{-2x}{(1-x^2)^2}\right)}{(e^{-\frac{1}{x^2}}+e^{-\frac{1}{1-x^2}})^2}\ge0.$$

Proof. Simplifying, it is enough to show that

$$\frac{2}{x^3}\left(e^{-\frac{1}{x^2}}+e^{-\frac{1}{1-x^2}}\right)+e^{-\frac{1}{x^2}}\cdot\frac{2}{x^3}+e^{-\frac{1}{1-x^2}}\cdot\frac{-2x}{(1-x^2)^2}\ge0.$$

If $x\le\frac{\sqrt{2}}{2}$, that is $x^2\le\frac{1}{2}$, then notice $\frac{1}{x^3}\ge\frac{x}{(1-x^2)^2}$, which proves the above. Thus from now on suppose $x\ge\frac{\sqrt{2}}{2}$. Rewrite the above desired inequality as

$$\begin{aligned}&\frac{1}{x^3}\left(2e^{-\frac{1}{x^2}}+e^{-\frac{1}{1-x^2}}\right)-\frac{x}{(1-x^2)^2}e^{-\frac{1}{1-x^2}}\ge0\\ \iff&(1-x^2)^2(2e^{-\frac{1}{x^2}}+e^{-\frac{1}{1-x^2}})\ge x^4e^{-\frac{1}{1-x^2}}\\ \iff&2(1-x^2)^2e^{-\frac{1}{x^2}}\ge(2x^2-1)e^{-\frac{1}{1-x^2}}\\ \iff&e^{\frac{1}{1-x^2}-\frac{1}{x^2}}\ge\frac{2x^2-1}{2(1-x^2)^2}.\end{aligned}$$

Notice $\frac{1}{1-x^2}-\frac{1}{x^2}\ge0$ since $2x^2\ge1$, thus by series expansion, it suffices to show

$$1+\frac{1}{1-x^2}-\frac{1}{x^2}+\frac{1}{2}\left(\frac{1}{1-x^2}-\frac{1}{x^2}\right)^2+\frac{1}{6}\left(\frac{1}{1-x^2}-\frac{1}{x^2}\right)^3\ge\frac{2x^2-1}{2(1-x^2)^2}.$$

Explicitly expanding the above, because $0\le x\le1$, this is equivalent to the inequality

$$6x^6\left(1-x^2\right)^3+6x^4\left(1-x^2\right)^2\left(2x^2-1\right)+\left(2x^2-1\right)^3+3\left(2x^2-1\right)^2x^2\left(1-x^2\right)-3x^6\left(1-x^2\right)\left(2x^2-1\right)\ge0$$

for $x\in[\frac{\sqrt{2}}{2},1]$. Replacing x^2 by x, the left hand side of the above expands to

$$h(x)=-6x^6+36x^5-69x^4+65x^3-33x^2+9x-1.$$

We want to show $h(x) \geq 0$ for $x \in \left[\frac{1}{2}, 1\right]$. This can be directly checked by computer, but we also give a proof by hand. Noting $h(\frac{1}{2}), h'(\frac{1}{2}), h''(\frac{1}{2}) \geq 0$, it is enough to show $h'''(x) \geq 0$ on $[\frac{1}{2}, 1]$, or equivalently

$$h_3(x) := -120x^3 + 360x^2 - 276x + 65 \geq 0 \forall x \in [\frac{1}{2}, 1].$$

However differentiating and applying the quadratic formula, we can check $h_3(x)$ attains a minimum value on $\left[\frac{1}{2}, 1\right]$ at $x = 1 - \sqrt{\frac{7}{30}} \approx 0.517$, and that this minimum value is strictly positive. This completes the proof. ■