Connections Between Gradient Based Optimization, Sampling and Lyapunov Functions

August Chen

Ayush Sekhari

Gradient Based Optimization

Goal:

$$\text{Minimize} : F(\mathbf{x})$$

- Deterministic optimization, we get access to $\nabla F(\mathbf{x})$
- Learning or Stochastic optimization: $F(\mathbf{x}) = \mathbb{E}_{z \sim D}[f(\mathbf{x}; z)]$
- We get access to stochastic gradients of form $\nabla f(\mathbf{x}; z_t)$ where $z_t \sim D$

$$\mathbb{E}_{z_t \sim D}[\nabla f(\mathbf{x}; z_t)] = \nabla F(\mathbf{x})$$

Score Function Based Sampling

Goal:

Sample from distribution with density : $p(\mathbf{x}) = e^{-\beta F(\mathbf{x})}/Z_\beta$

- We get access to "score function" (gradient): $\nabla F(\mathbf{x})$

Gradient Based Optimization Vs Sampling

Optimization

Minimize $F(\mathbf{x})$

Sampling

Sample from $p(\mathbf{x}) \propto \exp(-\beta F(\mathbf{x}))$

Gradient Based Optimization Vs Sampling

Optimization

$$\text{Minimize} \quad F(\mathbf{x})$$

Sampling

$$\text{Sample from } p(\mathbf{x}) \propto \exp(-\beta F(\mathbf{x}))$$

Gradient Descent (GD):

$$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta \nabla F(\mathbf{x}_{t-1})$$

Gradient Based Optimization Vs Sampling

Optimization

Sampling

$$\text{Minimize} \quad F(\mathbf{x})$$

$$\text{Sample from } p(\mathbf{x}) \propto \exp(-\beta F(\mathbf{x}))$$

Gradient Descent (GD):

$$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta \nabla F(\mathbf{x}_{t-1})$$

Stochastic GD (learning):

$$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta \nabla f(\mathbf{x}_{t-1}, z_t)$$

Gradient Based Optimization Vs Sampling

Optimization

$$\text{Minimize} \quad F(\mathbf{x})$$

Sampling

$$\text{Sample from } p(\mathbf{x}) \propto \exp(-\beta F(\mathbf{x}))$$

Gradient Descent (GD):

$$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta \nabla F(\mathbf{x}_{t-1})$$

Stochastic GD (learning):

$$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta \nabla f(\mathbf{x}_{t-1}, z_t)$$

Gradient Langevin Dynamics (GLD)

$$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta \nabla F(\mathbf{x}_{t-1}) + \sqrt{2\eta\beta^{-1}}\epsilon_t$$

Gradient Based Optimization Vs Sampling

Optimization

$$\text{Minimize} \quad F(\mathbf{x})$$

Sampling

$$\text{Sample from } p(\mathbf{x}) \propto \exp(-\beta F(\mathbf{x}))$$

Gradient Descent (GD):

$$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta \nabla F(\mathbf{x}_{t-1})$$

Stochastic GD (learning):

$$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta \nabla f(\mathbf{x}_{t-1}, z_t)$$

Gradient Langevin Dynamics (GLD)

$$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta \nabla F(\mathbf{x}_{t-1}) + \sqrt{2\eta\beta^{-1}}\epsilon_t$$

Stochastic GLD (learning)

$$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta \nabla F(\mathbf{x}_{t-1}) + \sqrt{2\eta\beta^{-1}}\epsilon_t$$

Gradient Based Optimization Vs Sampling

Optimization

$$\text{Minimize} \quad F(\mathbf{x})$$

Gradient Descent (GD):

$$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta \nabla F(\mathbf{x}_{t-1})$$

Stochastic GD (learning):

$$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta \nabla f(\mathbf{x}_{t-1}, z_t)$$

Gradient Langevin Dynamics (GLD)

$$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta \nabla F(\mathbf{x}_{t-1}) + \sqrt{2\eta\beta^{-1}}\epsilon_t$$

Stochastic GLD (learning)

$$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta \nabla F(\mathbf{x}_{t-1}) + \sqrt{2\eta\beta^{-1}}\epsilon_t$$

Sampling

$$\text{Sample from } p(\mathbf{x}) \propto \exp(-\beta F(\mathbf{x}))$$

Langevin Monte Carlo Sampling:

$$\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta \nabla F(\mathbf{x}_{t-1}) + \sqrt{2\eta\beta^{-1}}\epsilon_t$$

Questions

QUESTIONS

When do these algorithms work?

Questions

When do these algorithms work?

Whats the relationship between Gradient based Sampling and Optimization?

Questions

When do these algorithms work?

Whats the relationship between Gradient based Sampling and Optimization?

Is there a unifying analysis technique?

Outline

1. Reviewing the continuous time (idealized) processes

GRADIENT FLOW

Consider the continuous time (idealized) Gradient Descent process:

$$d\mathbf{x}(t) = -\nabla F(\mathbf{x}(t))dt$$

- Think of $\mathbf{x}(0) = \mathbf{x}_0$ as the starting point
- w.l.o.g. assume F is minimized at $\mathbf{0}$ and that $F(\mathbf{0}) = 0$
- In general if we SGD or GD with some step size scheme could work, then we would expect this idealized process to work
- Eg. Given $\epsilon > 0$, and any starting point $\mathbf{x}_0$, there exists $t < \infty$ such that $F(\mathbf{x}(t)) \leq \epsilon$
- Define $\tau_\epsilon(\mathbf{x}_0)$ to be the smallest such time.

Langevin Diffusion Process

Consider the continuous time (idealized) Gradient Langevin Dynamics process:

$$d\mathbf{x}(t) = -\nabla F(\mathbf{x}(t))dt + \sqrt{2\beta^{-1}}d\mathbf{B}(t)$$

where $\mathbf{B}(t)$ is the standard brownian motion in $\mathbb{R}^d$

- In general if we assume SGLD or GLD would works, would expect this idealized process to work
- Define Hitting time $\tau_\epsilon(\mathbf{x}_0) = \inf\{t : F(\mathbf{x}(t)) \leq \epsilon\}$.
- We would expect hitting time to be well behaved

Generator for a Markov Process

Definition

The (infinitesimal) generator of a Markov process $\mathbf{x}(t)$ is the operator $\mathcal{L}$ defined on all (sufficiently differentiable) functions f by

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

- Gradient Flow: $\mathcal{L}^{GF}f(\mathbf{x}) = -\langle \nabla F(\mathbf{x}), \nabla f(\mathbf{x}) \rangle$
- Langevin Diffusion: $\mathcal{L}^{LD}f(\mathbf{x}) = -\langle \nabla F(\mathbf{x}), \nabla f(\mathbf{x}) \rangle + \beta^{-1} \Delta f(\mathbf{x})$ where Δ is the Laplacian operator

Outline

2. Lyapunov Functions

Lyapunov Potential

Definition

A non-negative function Φ is a Lyapunov Potential on open set $\mathcal{A}$ if $\Phi \geq 1$ and on set $\mathcal{A}$ we have:

$$-\mathcal{L}\Phi \geq \lambda\Phi$$

- For optimization we will consider the set $\mathcal{A} = \{\mathbf{x} \in \mathbb{R}^d : F(\mathbf{x}) > \epsilon\}$

Outline

3. Analysis of Optimization Using Lyapunov Function

Why Lyapunov Function Helps: GD

Say the Lyapunov potential was H-smooth and function F is L Lipschitz, then

$$\begin{aligned}
\Phi(\mathbf{x}_t) &= \Phi(\mathbf{x}_{t-1} - \eta \nabla F(\mathbf{x}_{t-1})) \\
&\le \Phi(\mathbf{x}_{t-1}) - \eta \langle \nabla F(\mathbf{x}_{t-1}), \nabla \Phi(\mathbf{x}_{t-1}) \rangle + \frac{H\eta^2}{2} \|\nabla F(\mathbf{x}_{t-1})\|_2^2 \\
&\le \Phi(\mathbf{x}_{t-1}) - \eta \langle \nabla F(\mathbf{x}_{t-1}), \nabla \Phi(\mathbf{x}_{t-1}) \rangle + \frac{HL^2\eta^2}{2} \\
&\le \Phi(\mathbf{x}_{t-1}) - \eta\lambda\Phi(\mathbf{x}_{t-1}) + \frac{HL^2\eta^2}{2}
\end{aligned}$$

Rearranging and taking an average:

$$1 \le \frac{1}{T}\sum_{t=1}^{T} \Phi(\mathbf{x}_{t-1}) \le \frac{\Phi(\mathbf{x}_0)}{\eta\lambda} + \frac{HL^2\eta}{2T}$$

Setting η, T cannot be too large before we get contradiction.

Lyapunov Function for GF Exists + is smooth

\+ **Variance of gradient estimates are bounded**

Gradient Descent Works

Stochastic Gradient Descent Learns

Smoothness of Potential can be replaced by more general Self-boundedness of Gradient Norm

Why Lyapunov Function Helps: GLD

Same idea: Taylor up to one higher order

$$
\begin{aligned}
\mathbb{E}_{\epsilon_t}[\Phi(\mathbf{x}_t)] &= \Phi(\mathbf{x}_{t-1}) - \eta\nabla F(\mathbf{x}_{t-1}) + \sqrt{\eta\beta^{-1}}\epsilon_t) \\
&\leq \Phi(\mathbf{x}_{t-1}) - \eta\left\langle \nabla F(\mathbf{x}_{t-1}), \nabla\Phi(\mathbf{x}_{t-1})\right\rangle \\
&\qquad + 4\eta\beta^{-1}\mathbb{E}_{\epsilon_t}\left[\epsilon_t^\top \nabla^2\Phi(\mathbf{x}_{t-1})\epsilon_t\right] + \frac{HL^2\eta^2}{2} + \text{higher order} \\
&= \leq \Phi(\mathbf{x}_{t-1}) - \eta\left\langle \nabla F(\mathbf{x}_{t-1}), \nabla\Phi(\mathbf{x}_{t-1})\right\rangle \\
&\qquad + \eta\beta^{-1}\Delta\Phi(\mathbf{x}_{t-1}) + \frac{HL^2\eta^2}{2} + \text{higher order} \\
&\leq \Phi(\mathbf{x}_{t-1}) - \eta\lambda\Phi(\mathbf{x}_{t-1}) + \frac{HL^2\eta^2}{2} + \text{higher order}
\end{aligned}
$$

Lyapunov Function for LD Exists + is higher order smoothness

+ **Variance of gradient estimates are bounded**

Gradient Langevin Dynamics works

SGLD Works for Learning

Lyapunov Function

Definition

A non-negative function Φ is a Lyapunov Potential on open set $\mathcal{A}$ if $\Phi \geq 1$ and on set $\mathcal{A}$ we have:

$$-\mathcal{L}\Phi \geq \lambda\Phi$$

- For optimization we will consider the set $\mathcal{A} = \{\mathbf{x} \in \mathbb{R}^d : F(\mathbf{x}) > \epsilon\}$
- Using [Cattiaux & Guillin '17] (for LD and GF just plain calculus): Existence of such potential Φ is equivalent to existence of $\theta > 0$ s.t.

$$\mathbb{E}[\exp(\theta\tau_{\mathcal{A}^c})] < \infty$$

In other words the continuous time process work (for both GF and GLD) if and only if such Lyapunov potentials exist.

Gradient Flow or Langevin Diffussion Works

$$\Updownarrow$$

A corresponding Lyapunov Function
on the ϵ sub-optimal set exists

Outline

4. Sampling From Isoperimetric Inequalities

Poincare Inequality

Definition

A measure μ on $\mathbb{R}^d$ satisfies Poincare Inequality (PI) with constant $C_{PI}(\mu)$ if for all infinitely differentiable functions f,

$$\mathrm{Var}_\mu(f) \leq C_{PI}(\mu) \int \|\nabla f\|^2 d\mu$$

- When Variance is replaced by entropy of f^2 the above inequality is referred to as Log-Sobolev Inequality (LSI) with constant $C_{LSI}(\mu)$
- Taking measure μ_β to be given by the density $p(\mathbf{x}) = e^{-\beta F(\mathbf{x})}/Z$, PI and LSI are properties on F.
- For a function F, μ_β satisfying PI is a much weaker condition than F being convex or PL or KL or pretty much most conditions under which GD and friends are shown to converge.

Isoperimetric Inequalities Implies Sampling

- Letting π_T be measure from SDE for time T and π_0 be initialization for LD:
$$\chi^2(\pi_T \| \mu_\beta) \leq e^{-2T/C_{PI}(\mu_\beta)} \chi^2(\pi_0 \| \mu_\beta).$$
- From existing literature, inequalities like PI and LSI imply that sampling is possible with upper bounds on rates of convergence
- Such isoperimetric inequalities are amongst the more general conditions under which we can derive sampling results
- [Cattiaux & Guillin '17]: Existence of Lyapunov function for Langevin Diffusion is equivalent to μ_β satisfying PI.

Outline

5. Isoperimetric Inequalities and Lyapunov Functions

Lyapunov Function for LD Exists

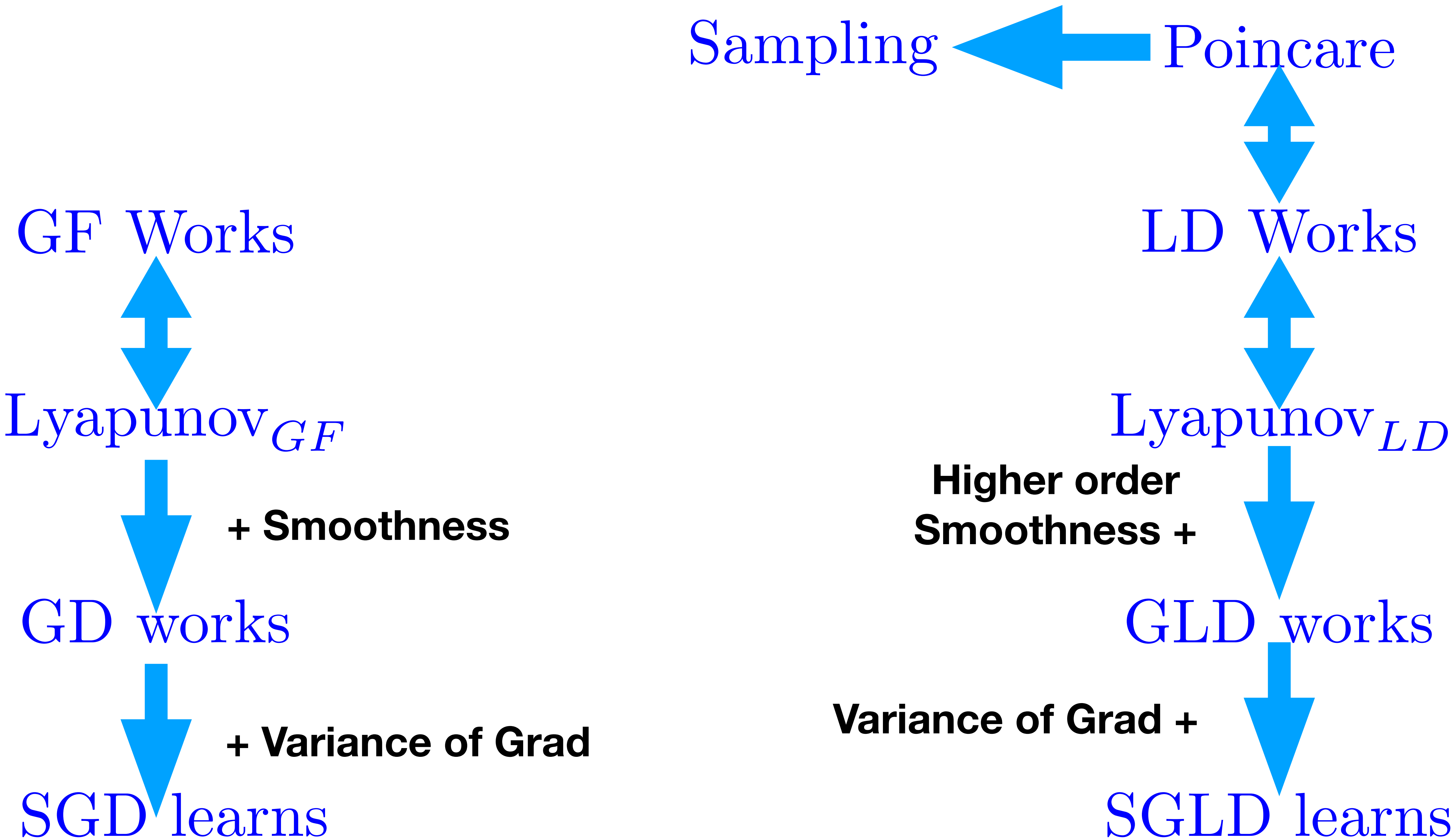
Poincare Inequality Holds

Sampling

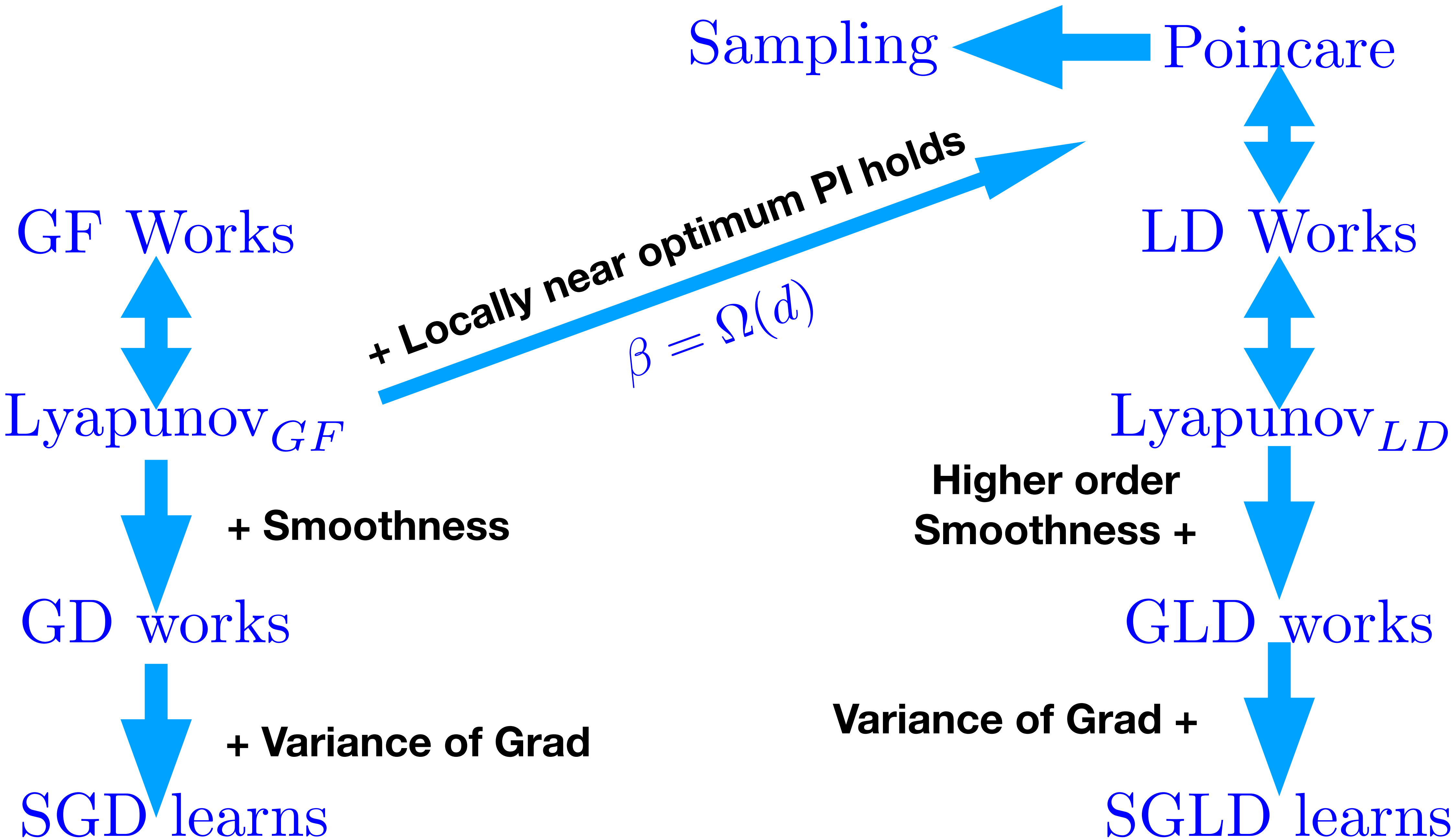
RESULTS

Problem Setting	Our Result	Best in Literature
GLD Poincaré & Lipschitz	$\widetilde{O}\left(\max\left\{d^3 \mathsf{C}_{\mathrm{PI}}(\mu_\beta)^3, \frac{d^2 \mathsf{C}_{\mathrm{PI}}(\mu_\beta)^2}{\varepsilon^2}\right\}\right)$	$\widetilde{O}\left(\frac{d^{14} \mathsf{C}_{\mathrm{PI}}(\mu_\beta)^3}{\varepsilon^{16}}\right)$ (Balasubramanian et al., 2022)
SGLD Poincaré & Lipschitz	$\widetilde{O}\left(\max\left\{d^3 \mathsf{C}_{\mathrm{PI}}(\mu_\beta)^3, \frac{d^2 \mathsf{C}_{\mathrm{PI}}(\mu_\beta)^2}{\varepsilon^2}\right\}\right)$	No finite guarantee
SGLD smooth & dissipative	$\widetilde{O}\left(\max\left\{d^3 \mathsf{C}_{\mathrm{PI}}(\mu_\beta)^3, \frac{d^2 \mathsf{C}_{\mathrm{PI}}(\mu_\beta)^2}{\varepsilon^2}\right\}\right)$	$\widetilde{O}\left(\min\left\{\frac{d^8 \mathsf{C}_{\mathrm{PI}}(\mu_\beta)^2}{\varepsilon^4}, \frac{d^7}{\varepsilon^5 \lambda_*^5}\right\}\right)$ (Xu et al., 2018; Zou et al., 2021)

Putting it All Together



Putting it All Together



Assumptions

Locally PI Assumption: For small enough $l > 0$ there exists radius $r(l) > 0$ s.t. $\{x : F(x) \leq l\} \subset B_2(\mathcal{X}^*, r(l))$ s.t. the measure $\mu_{\beta,\text{Local}(l)}$ satisfies Poincare Inequality with constant $C_{\text{PI,Local}}(l)$. Here $\mathcal{X}^*$ is the set of minima of F and $B_2(\mathcal{X}^*, r(l)) = \{x : d(x, X^*) \leq r(l)\}$.

Dissipativity: $c_1, c_2, R > 0$ s.t. for some $x^* \in \mathcal{X}^*$, we have that $\forall x \in B(x^*, R)^c$,

$$\langle \nabla F(x), x - x^* \rangle \geq c_1 F(x) \quad \text{and} \quad F(x) \geq c_2 \|x - x^*\|$$

Note: Above condition is more general than dissipativity

Optimizability with GF Implies Poincare

Theorem

When $\beta = \Omega(d)$, *under the assumptions that* μ_β *is Locally PI and the dissipativity assumption, we have that if* F *is optimizable using gradient flow, then the measure* μ_β *satisfies Poincare inequality with*

$$C_{PI}(\mu_\beta) = O\left(C_{\mathrm{PI,Local}} + \frac{1}{\beta}\right)$$

Remark: Under weak convexity + Quadratic tail growth of F Log sobolev Inequality also holds.

IMPLICATIONS

- Obtain Isoperimetric inequalities with $\text{poly}(d, 1/\beta)$ for a host of non-log-concave measures. Eg. when F satisfies PL, KL conditions or is quasar convex etc.

- Implies continuous time sampling result in TV for such measures under arbitrary initialization

- Under additional smoothness of potential we can obtain discrete time Langevin Monte Carlo algorithm with $\text{poly}(d, 1/\beta, 1/\epsilon)$ rates.

Weak Poincare Inequality

- Often we may not have convergence of GF from everywhere but only from set of initializations $\mathcal{S}$ (good set of initializations).
- In this case one can obtain a weaker notion of Poincare like inequality termed weak Poincare inequality.
- Under weak PI while mixing from arbitrary starting distribution may not work but appropriate warm start still works.

Definition

A measure μ on $\mathbb{R}^d$ is said to satisfy a $(C_{WPI}(\mu), \delta)$ Weak Poincare Inequality (PI) if for all infinitely differentiable f's, ($\mathrm{osc}(f) = \sup f - \inf f$)

$$\mathrm{Var}_\mu(f) \leq C_{WPI}(\mu) \int \|\nabla f\|^2 d\mu + \delta\, \mathrm{osc}(f)^2$$

Initialization Dependent GF to Weak Poincare

Theorem

When $\beta = \Omega(d)$*, under the assumptions that* μ_β *is Locally PI and the dissipativity assumption, we have that if* F *is optimizable using gradient flow when starting from a good initialization set* $\mathcal{S}$*, then the measure* μ_β *satisfies* $(C_{WPI}(\mu), \delta)$*Weak Poincare inequality with*

$$C_{PI}(\mu_\beta) = O\left(C_{\text{PI,Local}} + \frac{1}{\beta}\right), \quad \delta = O(\mu_\beta(S^c))$$

Summary

- Strong connection between optimizability using gradient flow and Isoperimetric inequalities
- Layman terms: Isoperimetric inequalities implies optimizability with gradient descent
- Layman terms: Optimizability using gradient descent implies sampling up to $\Omega(d)$ temperature regimes
- Implication: General conditions for GD to work like KL, PL, quasar convex, linearizability imply sampling using objective as energy function for appropriate temp

Thanks!