Stochasticity of Deterministic Gradient Descent: Quantitative Local Min Escape in Multiscale Landscape

Molei Tao  School of Mathematics, Georgia Tech, USA

November 12, 2023  Caltech ACM
Machine Learning & Applied Computational Math
Machine Learning & Applied Computational Math

→ ML for problems in computing, sciences & engineering
Machine Learning & Applied Computational Math

→ ML for problems in computing, sciences & engineering

← how ACM helps design and analyze optimization, sampling, and deep learning practices
Machine Learning & Applied Computational Math

→ ML for problems in computing, sciences & engineering

← how ACM helps design and analyze optimization, sampling, and deep learning practices

common belief: **large learning rate** is good
common belief: \textbf{large learning rate is good}

\[
\min f(x) \quad x_{k+1} = x_k - h \nabla f(x_k)
\]

Machine Learning & Applied Computational Math

→ ML for problems in computing, sciences & engineering

← how ACM helps design and \textit{analyze}
  optimization, sampling, and \textit{deep learning practices}
Machine Learning & Applied Computational Math

→ ML for problems in computing, sciences & engineering

← how ACM helps design and **analyze**
optimization, sampling, and **deep learning practices**

**common belief:** **large learning rate** is **good**

1\textsuperscript{st} thought:

\[
\min f(x) \quad x_{k+1} = x_k - h \nabla f(x_k)
\]

Gradient Descent

\[
\begin{align*}
\hat{x}_i & \approx x(ih) \\
\dot{x} & = -\nabla f(x)
\end{align*}
\]

Gradient Flow
common belief: large learning rate is good

1st thought:

\[ \min f(x) \quad x_{k+1} = x_k - h \nabla f(x_k) \]

Gradient Descent

\[ x(T) \text{ with fixed } T \]
\[ \downarrow \]
\[ [T/h] \text{ steps} \]
\[ \downarrow \]
\[ \text{large } h? \]

\[ x_i \approx x(ih) \]
\[ \dot{x} = -\nabla f(x) \]

Gradient Flow
common belief: large learning rate is good

1st thought:

\[
\min f(x) \quad \quad x_{k+1} = x_k - h \nabla f(x_k)
\]

Gradient Descent

\[x(T) \text{ with fixed } T \downarrow [T/h] \text{ steps} \downarrow \text{large } h?\]

\[
\dot{x} = - \nabla f(x)
\]

Gradient Flow

→ ML for problems in computing, sciences & engineering

← how ACM helps design and analyze optimization, sampling, and deep learning practices
common belief: large learning rate is good

not just faster, but implicitly bias toward (desirable) global structures
common belief: **large learning rate** is good

not just faster, but implicitly bias toward (desirable) global structures

→ better training & better test accuracies
\[ \min_x f(x) := \sum_i d(\text{output}_i, g(x, \text{input}_i)) \]

- **training data (known)**
- **model (e.g., neural network) parameters**
Training

\[
\min_x f(x) := \sum_i d(\text{output}_i, g(x, \text{input}_i))
\]

model (e.g., neural network) parameters

training data (known)

trapped in local min \iff suboptimal training accuracy
training data (known) → suboptimal training accuracy

\[
\min_x f(x) := \sum_i d(\text{output}_i, g(x, \text{input}_i))
\]

model (e.g., neural network) parameters

large h? this talk

trapped in local min ↔ suboptimal training accuracy
Training

\[ \min_x f(x) := \sum_i d(\text{output}_i, g(x, \text{input}_i)) \]

Large \( h \)? This talk

Trapped in local min \( \iff \) Suboptimal training accuracy

Testing

\[ d(\overbrace{\text{output}_j}^{\text{test data}}, g(x_{\text{trained}}, \overbrace{\text{input}_j}^{\text{test data}})) \]
training

\[ \min_x f(x) := \sum_i d(\text{output}_i, g(x, \text{input}_i)) \]

large h? this talk

trapped in local min \iff suboptimal training accuracy

test data

\[ d(\text{output}_j, g(x_{\text{trained}}, \text{input}_j)) \]

test data

generalization

model (e.g., neural network) parameters

training data (known)
training

$$\min_x f(x) := \sum_i d(output_i, g(x, input_i))$$

trapped in local min \iff suboptimal training accuracy

large h? this talk

large h? offline

generalization

test data

d(output_j, g(x_{trained}, input_j))

matters

model (e.g., neural network) parameters

training data (known)
Stochasticity of Deterministic Gradient Descent: Large Learning Rate for Multiscale Objective Function


Lingkai Kong$^1$ and Molei Tao$^1$

1 Georgia Institute of Technology (USA)
1-1. Introduction: GD with small and large LRs

$$\min f(x)$$

$$x_{k+1} = x_k - h \nabla f(x_k)$$

Gradient Descent
1-1. Introduction: GD with small and large LRs

$$\min f(x)$$

$$x_{k+1} = x_k - h \nabla f(x_k)$$

Gradient Descent
1-2. Introduction: GD with `large’ LR for multiscale objective -- phenomenology
1-2. Introduction: GD with `large’ LR for multiscale objective -- phenomenology
1-2. Introduction: GD with `large’ LR for multiscale objective -- phenomenology
1-2. Introduction: GD with `large’ LR for multiscale objective -- phenomenology
1-2. Introduction: GD with `large’ LR for multiscale objective -- phenomenology

deterministic dynamics (no stochastic gradient) but `stochastic’ behavior similar to SGD
2-1. Theory: the setup: multiscale objective function

- Multiscale objective + deterministic GD with large LR

  → Chaos

  → 'Stochastic' behaviors
multiscale objective + deterministic GD with large LR

multiscale objective fct. \( f : \mathbb{R}^d \to \mathbb{R} \). \( f(x) := f_0(x) + f_{1,\epsilon}(x) \)
multiscale objective + deterministic GD with large LR

 multiscale objective fct. \( f : \mathbb{R}^d \rightarrow \mathbb{R} \). \( f(x) := f_0(x) + f_{1,\varepsilon}(x) \)
2-1. Theory: the setup: multiscale objective function

**multiscale objective** +
deterministic GD with large LR

multiscale objective fct. $f : \mathbb{R}^d \rightarrow \mathbb{R}$. $f(x) := f_0(x) + f_{1,\epsilon}(x)$
2-1. Theory: the setup: multiscale objective function

**multiscale objective** + deterministic GD with large LR

multiscale objective fct. $f : \mathbb{R}^d \rightarrow \mathbb{R}$. $f(x) := f_0(x) + f_{1,\epsilon}(x)$

**EX** $f_{1,\epsilon} := \epsilon f_1 \left( \frac{x}{\epsilon} \right)$

$f_1$ periodic

![Graphs of $f_0$, $f$](image)
Multiscale objective + deterministic GD with large LR

 multiscale objective fct. $f : \mathbb{R}^d \to \mathbb{R}$. $f(x) := f_0(x) + f_{1,\epsilon}(x)$

EX $f_{1,\epsilon} := \epsilon f_1 \left( \frac{x}{\epsilon} \right)$

$f_1$ periodic

EX $f_{1,\epsilon} := \epsilon F_1 \left( \frac{\omega_1 x}{\epsilon}, \frac{\omega_2 x}{\epsilon}, \ldots, \frac{\omega_N x}{\epsilon} \right)$

$F_1$ 1-periodic in each argument

micro $\epsilon \ll 1$
multiscale objective +
deterministic GD with large LR

multiscale objective fct. $f : \mathbb{R}^d \to \mathbb{R}$. $f(x) := f_0(x) + f_{1,\epsilon}(x)$

EX $f_{1,\epsilon} := \epsilon f_1 \left( \frac{x}{\epsilon} \right)$

$f_1$ periodic

EX $f_{1,\epsilon} := \epsilon F_1 \left( \frac{\omega_1 x}{\epsilon}, \frac{\omega_2 x}{\epsilon}, \ldots, \frac{\omega_N x}{\epsilon} \right)$

$F_1$ 1-periodic in each argument

more general:

**Condition 1.** \( \exists \) bounded random variable \( \zeta \) with \( P(\zeta = 0) < 1 \), s.t. for any bounded rectangle \( \Gamma \subset \mathbb{R}^d \) with area \( |\Gamma| > 0 \), \( -\nabla f_{1,\epsilon}(U_\Gamma) \xrightarrow{w} \zeta \) when \( \epsilon \to 0 \). Assume without loss of generality that \( \mathbb{E}\zeta = 0 \) (nonzero mean can be absorbed into \( f_0 \)).

**Condition 2.** \( \epsilon \nabla^2 f_{1,\epsilon} \) is uniformly bounded as \( \epsilon \to 0 \), and \( \exists m \in \mathbb{R} \), s.t. for any bounded rectangle \( \Gamma \subset \mathbb{R}^d \) whose area \( |\Gamma| > 0 \), \( \mathbb{E} \left[ \ln \|\epsilon \nabla^2 f_{1,\epsilon}(U_\Gamma)\|_2 \right] \to m \).
2-1. Theory: the setup: multiscale objective function

multiscale objective + deterministic GD with large LR

multiscale objective fct. \( f : \mathbb{R}^d \to \mathbb{R} \). \( f(x) := f_0(x) + f_{1,\epsilon}(x) \)

EX \( f_{1,\epsilon} := \epsilon f_1 \left( \frac{x}{\epsilon} \right) \)
\( f_1 \) periodic

EX \( f_{1,\epsilon} := \epsilon F_1 \left( \frac{\omega_1 x}{\epsilon}, \frac{\omega_2 x}{\epsilon}, \ldots, \frac{\omega_N x}{\epsilon} \right) \)
\( F_1 \) 1-periodic in each argument

more general: O(1) 1st-derivative with a weak limit

**Condition 1.** \( \exists \) bounded random variable \( \zeta \) with \( P(\zeta = 0) < 1 \), s.t. for any bounded rectangle \( \Gamma \subset \mathbb{R}^d \) with area \( |\Gamma| > 0 \), \( -\nabla f_{1,\epsilon}(U_\Gamma) \xrightarrow{w} \zeta \) when \( \epsilon \to 0 \). Assume without loss of generality that \( \mathbb{E}\zeta = 0 \) (nonzero mean can be absorbed into \( f_0 \)).

**Condition 2.** \( \epsilon \nabla^2 f_{1,\epsilon} \) is uniformly bounded as \( \epsilon \to 0 \), and \( \exists m \in \mathbb{R} \), s.t. for any bounded rectangle \( \Gamma \subset \mathbb{R}^d \) whose area \( |\Gamma| > 0 \), \( \mathbb{E} \left[ \ln \|\epsilon \nabla^2 f_{1,\epsilon}(U_\Gamma)\|_2 \right] \to m. \)
2-1. Theory: the setup: multiscale objective function

multiscale objective +
deterministic GD with large LR

multiscale objective fct. \( f : \mathbb{R}^d \to \mathbb{R} \). \( f(x) := f_0(x) + f_{1,\epsilon}(x) \)

EX \[ f_{1,\epsilon} := \epsilon f_1 \left( \frac{x}{\epsilon} \right) \]

\( f_1 \) periodic

EX \[ f_{1,\epsilon} := \epsilon F_1 \left( \frac{x_1}{\epsilon}, \frac{x_2}{\epsilon}, \ldots, \frac{x_N}{\epsilon} \right) \]

\( F_1 \) 1-periodic in each argument

more general:

O(1) 1st-derivative with a weak limit

Condition 1. \( \exists \) bounded random variable \( \zeta \) with \( P(\zeta = 0) < 1 \), s.t. for any bounded rectangle \( \Gamma \subset \mathbb{R}^d \) with area \( |\Gamma| > 0 \), \( -\nabla f_{1,\epsilon}(U_\Gamma) \xrightarrow{w} \zeta \) when \( \epsilon \to 0 \). Assume without loss of generality that \( \mathbb{E}\zeta = 0 \) (nonzero mean can be absorbed into \( f_0 \)).

Condition 2. \( \epsilon \nabla^2 f_{1,\epsilon} \) is uniformly bounded as \( \epsilon \to 0 \), and \( \exists m \in \mathbb{R} \), s.t. for any bounded rectangle \( \Gamma \subset \mathbb{R}^d \) whose area \( |\Gamma| > 0 \), \( \mathbb{E} \left[ \ln \|\epsilon \nabla^2 f_{1,\epsilon}(U_\Gamma)\|_2 \right] \to m \).

O(1/\epsilon) 2nd-derivative with a limiting expection
multiscale objective +
deterministic GD with large LR

\[ f(x) := f_0(x) + f_{1,\epsilon}(x) \]
\[ \text{e.g., } f_{1,\epsilon} = \epsilon f_1(x/\epsilon) \]

\[ x \mapsto x - \eta \nabla f(x), \quad \text{i.e.} \quad x \mapsto x - \eta (\nabla f_0(x) + \nabla f_{1,\epsilon}(x)) \]
multiscale objective +
deterministic GD with large LR

\[ f(x) := f_0(x) + f_{1,\epsilon}(x) \]
\[ \text{e.g., } f_{1,\epsilon} = \epsilon f_1(x/\epsilon) \]

\[ x \mapsto x - \eta \nabla f(x), \quad \text{i.e.} \quad x \mapsto x - \eta (\nabla f_0(x) + \nabla f_{1,\epsilon}(x)) \]

small LR?
multiscale objective + deterministic GD with **large** LR

\[ f(x) := f_0(x) + f_{1,\epsilon}(x) \]

\[ e.g., f_{1,\epsilon} = \epsilon f_1(x/\epsilon) \]

\[ x \mapsto x - \eta \nabla f(x), \quad \text{i.e.} \quad x \mapsto x - \eta(\nabla f_0(x) + \nabla f_{1,\epsilon}(x)) \]

small LR? \[ \eta \ll 1/L \]

i.e., \[ \eta \ll \epsilon \]

\[ L: \text{Lipschitz const. of} \ \nabla f, \ L = \mathcal{O}(1/\epsilon) \]

resolve the small scale \(\rightarrow\) conv. to local min
multiscale objective +
deterministic GD

\[ f(x) := f_0(x) + f_{1,\epsilon}(x) \]
e.g., \( f_{1,\epsilon} = \epsilon f_1(x/\epsilon) \)

\[ x \mapsto x - \eta \nabla f(x), \quad \text{i.e.} \quad x \mapsto x - \eta (\nabla f_0(x) + \nabla f_{1,\epsilon}(x)) \]

small LR? \( \eta \ll 1/L \)

\( L: \) Lipschitz const. of \( \nabla f \), \( L = O(1/\epsilon) \)

i.e., \( \eta \ll \epsilon \)

resolve the small scale \( \rightarrow \) conv. to local min

bad LR? \( \eta \gg 1 \)

can’t even resolve the large scale

i.e., \( x \mapsto x - \eta \nabla f_0(x) \) \( \text{is unstable} \)
multiscale objective + deterministic GD with large LR

\[ f(x) := f_0(x) + f_{1,\epsilon}(x) \]
e.g., \( f_{1,\epsilon} = \epsilon f_1(x/\epsilon) \)

\[ x \mapsto x - \eta \nabla f(x), \quad \text{i.e.} \quad x \mapsto x - \eta (\nabla f_0(x) + \nabla f_{1,\epsilon}(x)) \]

small LR? \eta \ll 1/L \quad \text{L: Lipschitz const. of } \nabla f, \quad L = \mathcal{O}(1/\epsilon) \\
i.e., \eta \ll \epsilon \quad \text{resolve the small scale } \rightarrow \text{ conv. to local min}

bad LR? \eta \gg 1 \quad \text{can’t even resolve the large scale} \\
i.e., x \mapsto x - \eta \nabla f_0(x) \quad \text{is unstable}

in-between...

\[ \eta \ll \epsilon \quad \eta \sim \epsilon \quad \eta \gg \epsilon \quad \eta \gg \gg 1 \]

0 Converge to local minimum Instability
**2-1. Theory: the setup: large Learning Rate**

**multiscale objective + deterministic GD with large LR**

\[ f(x) := f_0(x) + f_{1,\epsilon}(x) \]

e.g., \( f_{1,\epsilon} = \epsilon f_1(x/\epsilon) \)

\[ x \mapsto x - \eta \nabla f(x), \quad \text{i.e.} \quad x \mapsto x - \eta (\nabla f_0(x) + \nabla f_{1,\epsilon}(x)) \]

small LR? \( \eta \ll 1/L \quad L: \text{Lipschitz const. of } \nabla f, \ L = \mathcal{O}(1/\epsilon) \)

i.e., \( \eta \ll \epsilon \quad \text{resolve the small scale } \rightarrow \text{conv. to local min} \)

bad LR? \( \eta \gg 1 \quad \text{can’t even resolve the large scale} \)

i.e., \( x \mapsto x - \eta \nabla f_0(x) \quad \text{is unstable} \)

in-between...

\[ \eta \ll \epsilon \quad \eta \approx \epsilon \quad \eta \gg \epsilon \quad \eta \gg 1 \]

- \( \eta \ll \epsilon \) Converge to local minimum
- \( \eta \approx \epsilon \) Instability
multiscale objective + deterministic GD with large LR

\[ f(x) := f_0(x) + f_{1, \epsilon}(x) \]
\[
eq f_0(x) + \epsilon f_1(x/\epsilon)
\]

\[ x \mapsto x - \eta \nabla f(x), \quad \text{i.e.} \quad x \mapsto x - \eta(\nabla f_0(x) + \nabla f_{1, \epsilon}(x)) \]

small LR? \quad \eta \ll 1/L \quad L: \text{Lipschitz const. of } \nabla f, \quad L = O(1/\epsilon)

\[
\text{i.e., } \eta \ll \epsilon
\]
\[
\text{resolve the small scale } \rightarrow \text{ conv. to local min}
\]

bad LR? \quad \eta \gg 1 \quad \text{can’t even resolve the large scale}

\[
\text{i.e., } x \mapsto x - \eta \nabla f_0(x) \quad \text{is unstable}
\]

in-between... large LR

\[
\eta \ll \epsilon \quad \eta \sim \epsilon \quad \eta \gg \epsilon
\]

0

Converge to local minimum

Instability
multiscale objective + deterministic GD with large LR

\[ f(x) := f_0(x) + f_{1,\epsilon}(x) \]
e.g., \( f_{1,\epsilon} = \epsilon f_1(x/\epsilon) \)

\[ x \mapsto x - \eta \nabla f(x), \quad \text{i.e.} \quad x \mapsto x - \eta (\nabla f_0(x) + \nabla f_{1,\epsilon}(x)) \]

small LR? \( \eta \ll 1/L \)

\( L: \) Lipschitz const. of \( \nabla f \), \( L = O(1/\epsilon) \)

i.e., \( \eta \ll \epsilon \)

resolve the small scale \( \rightarrow \) conv. to local min

bad LR? \( \eta \gg 1 \)

can’t even resolve the large scale

i.e., \( x \mapsto x - \eta \nabla f_0(x) \) is unstable

in-between... large LR

\[ \eta \ll \epsilon \]

Converge to local minimum

Local chaos

Global chaos

Instability
2-2. Theory: the local chaos regime

deterministic GD map

\[ \varphi(x) := x - \eta(\nabla f_0(x) + \nabla f_{1,\epsilon}(x)) \]
deterministic GD map

\[ \varphi(x) := x - \eta \left( \nabla f_0(x) + \nabla f_{1,\epsilon}(x) \right) \]

transition into chaos via period doublings

2-2. Theory: the local chaos regime

\[ \eta\ll\epsilon \quad \eta \sim \epsilon \quad \eta\gg\epsilon \quad \eta\gg1 \]

Converge to local minimum Local chaos Global chaos Instability
2-2. Theory: the local chaos regime

deterministic GD map

\[ \varphi(x) := x - \eta (\nabla f_0(x) + \nabla f_{1,\epsilon}(x)) \]

transition into chaos via period doublings

x stays within a microscopic potential well

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ O(\epsilon) \]
deterministic GD map

\[ \varphi(x) := x - \eta(\nabla f_0(x) + \nabla f_{1,\epsilon}(x)) \]

2-2. Theory: the local chaos regime

unimodal map within a well

transition into chaos via period doublings

x stays within a microscopic potential well

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ \eta \ll \epsilon \quad \eta \sim \epsilon \quad \eta \gg \epsilon \quad \eta \gg 1 \]

Converge to local minimum Local chaos Global chaos Instability
2-2. Theory: the local chaos regime

Deterministic GD map

\[ \varphi(x) := x - \eta(\nabla f_0(x) + \nabla f_{1,\epsilon}(x)) \]

Unimodal map within a well

Transition into chaos via period doublings

\( x \) stays within a microscopic potential well

\[ O(\epsilon) \]

\( \eta \ll \epsilon \)

Converge to local minimum

\( \eta \sim \epsilon \)

Local chaos

\( \eta \gg \epsilon \)

Global chaos

\( \eta \gg 1 \)

Instability
deterministic GD map

\[ \varphi(x) := x - \eta(\nabla f_0(x) + \nabla f_{1,\epsilon}(x)) \]

2-2. Theory: the local chaos regime

transition into chaos via period doublings

unimodal map within a well

x stays within a microscopic potential well

\[ O(\epsilon) \]

\[ O(\epsilon) \]

\[ \eta << \epsilon \]
Converge to local minimum

\[ \eta \sim \epsilon \]
Local chaos

\[ \eta >> \epsilon \]
Global chaos

\[ \eta >> 1 \]
Instability
deterministic GD map

\[ \varphi(x) := x - \eta(\nabla f_0(x) + \nabla f_{1,\epsilon}(x)) \]

\( \eta \) further increases (i.e. the \( \epsilon \to 0 \) regime)
deterministic GD map \( \varphi(x) := x - \eta(\nabla f_0(x) + \nabla f_{1,\epsilon}(x)) \)

\( \eta \) further increases (i.e. the \( \epsilon \to 0 \) regime)

- \( f_0 \) strongly convex, \( C^2 \), \( L \)-smooth

limiting distribution of \( \varphi \) iterates \( \approx \) nearly Gibbs

\[
\frac{1}{\mathcal{Z}} \exp \left( -\frac{2f_0(x)}{\eta\sigma^2} \right) dx + O(\eta^2)
\]

\( \eta \ll \epsilon \) Converge to local minimum

\( \eta \sim \epsilon \) Local chaos

\( \eta \gg \epsilon \) Global chaos

\( \eta \gg 1 \) Instability
2-3. Theory: the global chaos regime

Deterministic GD map

\[
\varphi(x) := x - \eta \left( \nabla f_0(x) + \nabla f_{1,\epsilon}(x) \right)
\]

\(\eta\) further increases \(\text{(i.e. the} \ \epsilon \to 0 \ \text{regime)}\) \quad \text{limiting distribution: nearly Gibbs}

\(f_0\) strongly convex)

\[
\frac{1}{Z} \exp \left( - \frac{2f_0(x)}{\eta \sigma^2} \right) dx + \mathcal{O}(\eta^2)
\]
deterministic GD map

\[ \varphi(x) := x - \eta(\nabla f_0(x) + \nabla f_{1,\epsilon}(x)) \]

\( \eta \) further increases (i.e. the \( \epsilon \to 0 \) regime)

limiting distribution: nearly Gibbs

(f_0 strongly convex)

\[ \frac{1}{Z} \exp \left( - \frac{2f_0(x)}{\eta \sigma^2} \right) dx + \mathcal{O}(\eta^2) \]
2-3. Theory: the global chaos regime

deterministic GD map

\[ \varphi(x) := x - \eta(\nabla f_0(x) + \nabla f_{1,\epsilon}(x)) \]

\( \eta \) further increases (i.e. the \( \epsilon \to 0 \) regime)

limiting distribution: nearly Gibbs

\( f_0 \) strongly convex

\[ \frac{1}{Z} \exp \left( -\frac{2f_0(x)}{\eta \sigma^2} \right) dx + O(\eta^2) \]

escape from local min, via large LR, without noise!
deterministic GD map

\[ \varphi(x) := x - \eta(\nabla f_0(x) + \nabla f_{1,\epsilon}(x)) \]

previously discussed (\( f_0 \) strongly convex):

\[ \frac{1}{Z} \exp \left( - \frac{2f_0(x)}{\eta \sigma^2} \right) dx + \mathcal{O}(\eta^2) \]

? non-convex \( f_0 \)
3. A practical example

Is f for real problems multiscale?
3. A practical example

5-16-2 FF neural network, regression, UCI Airfoil Self-Noise dataset, large LR
3. A practical example

EX  5-16-2 FF neural network, regression, UCI Airfoil Self-Noise dataset, large LR

Theory  multiscale data (mean + fluctuation) $\rightarrow$ multiscale loss?
3. A practical example

EX 5-16-2 FF neural network, regression, UCI Airfoil Self-Noise dataset, large LR

Theory

multiscale data (mean + fluctuation) → multiscale loss?

2 layer, periodic activation
chaotic dynamics can help deterministic GD escape microscopic local minima as it optimizes the loss samples a distribution
chaotic dynamics can help **deterministic** GD escape **microscopic** local minima as it optimizes the loss samples a distribution

?  why this matters
chaotic dynamics can help deterministic GD escape microscopic local minima as it optimizes the loss samples a distribution

? why this matters

deeper minimum $\rightarrow$ better training (& thus test) accuracies
Thank you for your attention and feedback!

Support:

NSF DMS-1847802, ECCS-1936776
GT Cullen-Peck Scholarship
Emory-GT AI.Humanity Award