Deep Relaxation: PDEs for optimizing Deep Neural Networks

Optimal Transport meets
Probability, Statistics and Machine Learning
BIRS-CMO
May 2, 2017

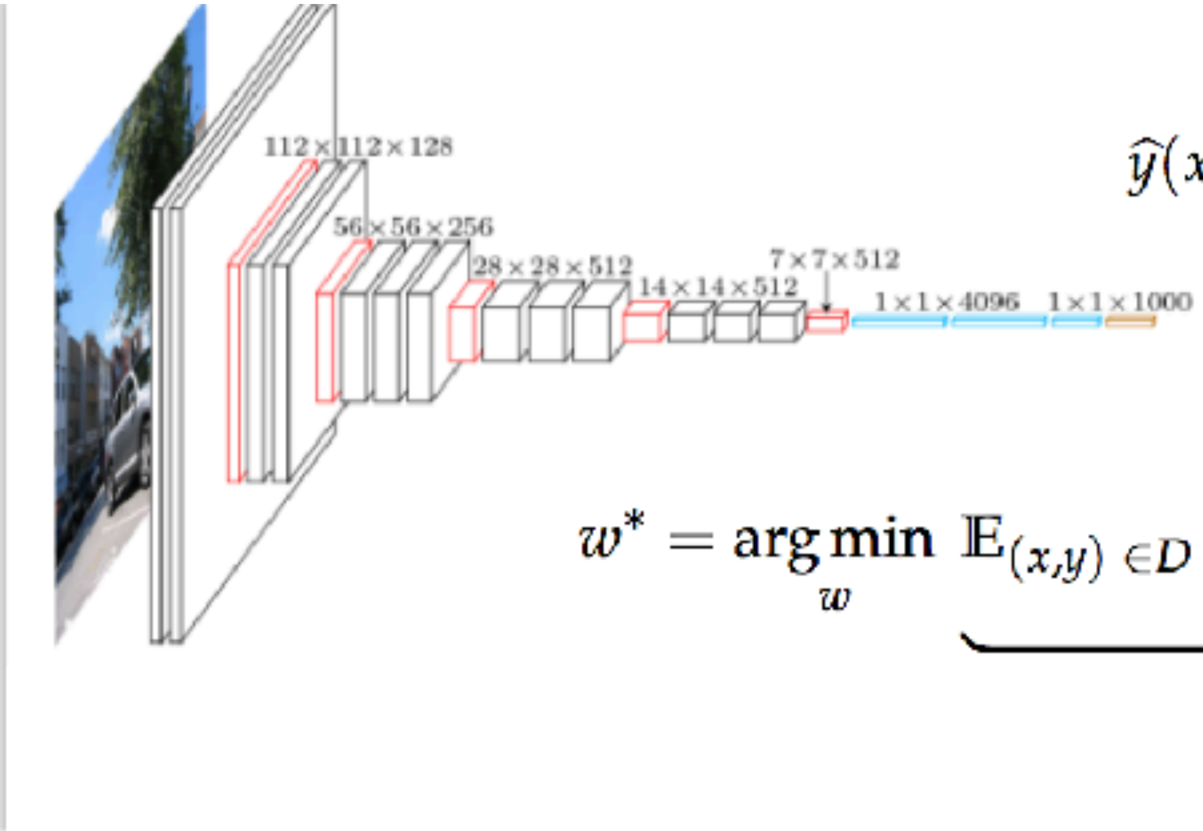
Adam Oberman (McGill)
Pratik Chaudhari, Stanley Osher, Stefano Soatto (UCLA)
Guillaume Carlier (CEREMADE)

Thanks to the Simons Foundation (grant 395980) and the hospitality of the UCLA math department

Outline

- 1. Introduction and Background
- 2. Math theory
- 3. Algorithm and efficiency
- 4. Computational Results

Deep Learning Background



$$\widehat{y}(x;\, w) = \sigma(w^p\, \sigma(w^{p-1}\, (\ldots \sigma(w^1\, x))\ldots))$$

$$w^* = \arg\min_w \underbrace{\mathbb{E}_{(x,y)\,\in D} \sum_{i=1}^{k} -y_i\, \log \widehat{y}_i(x;\, w)}_{\triangleq f(w)}$$

- Supervised Learning: training a deep network for image classification

- Goal: minimize the empirical loss function.

- Normally x data, w weight. Sequel, write f(x), x weights and suppress data.

FIGURE 2. MNIST

FIGURE 3. CIFAR-10

Related Recent Work

Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes

Carlo Baldassi[a,b,1], Christian Borgs[c], Jennifer T. Chayes[c], Alessandro Ingrosso[a,b], Carlo Lucibello[a,b], Luca Saglietti[a,b], and Riccardo Zecchina[a,b,d]

PNAS 2016

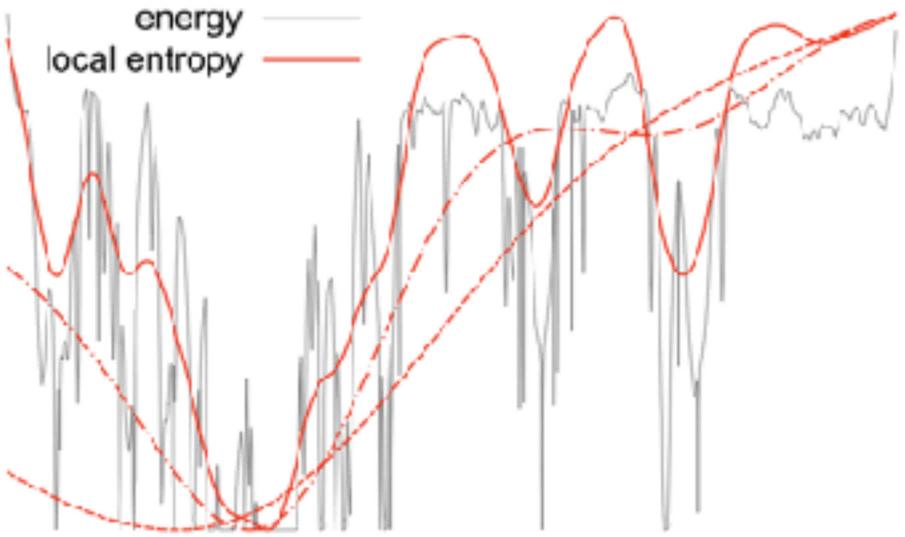
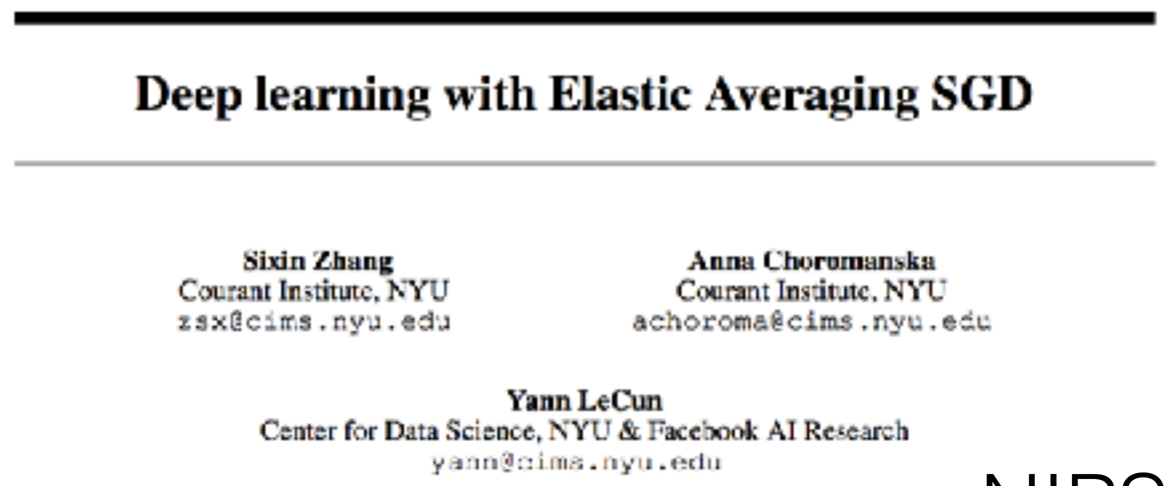


Fig. 1. Energy landscape compared with local entropy landscape in an illustrative toy example. The energy landscape (gray curve) can be very rugged, with a large number of narrow local minima. Some isolated global minima can also be observed on the right. On the left, there is a region of denser minima that coalesce into a wide global optimum. The red curves show the local entropy landscape (Eq. 2 with the opposite sign) computed at increasing values of the interaction parameter γ (i.e., at progressively finer scales).



Deep learning with Elastic Averaging SGD

Sixin Zhang
Courant Institute, NYU
zsx@cims.nyu.edu

Anna Choromanska
Courant Institute, NYU
achoroma@cims.nyu.edu

Yann LeCun
Center for Data Science, NYU & Facebook AI Research
yann@cims.nyu.edu

NIPS 2015

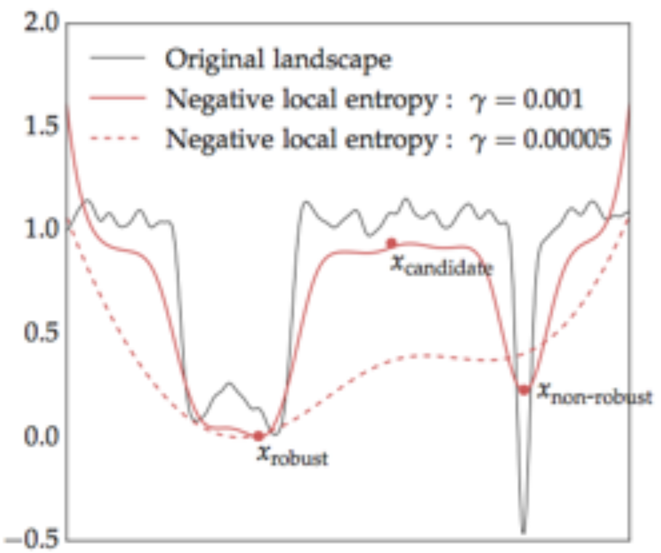
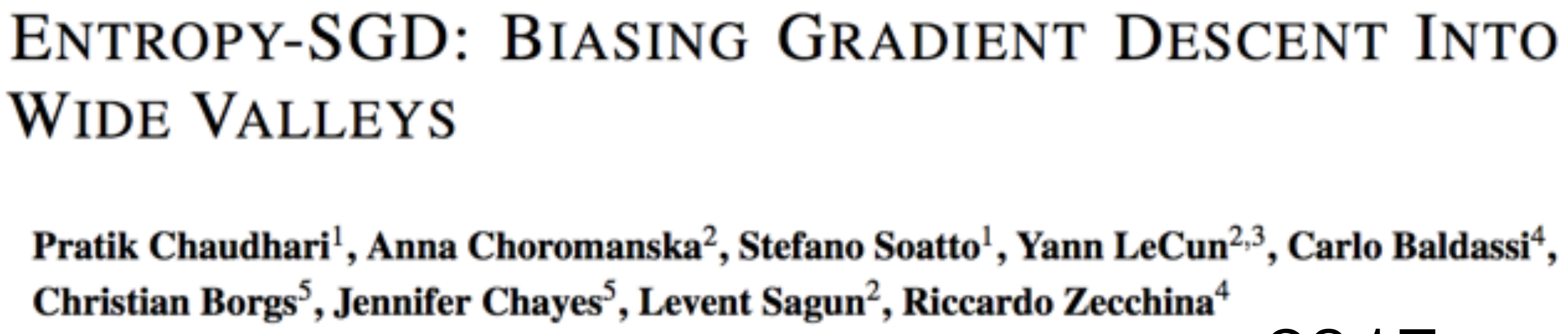


Figure 2: Local entropy concentrates on wide valleys in the energy landscape.



Entropy-SGD: Biasing Gradient Descent Into Wide Valleys

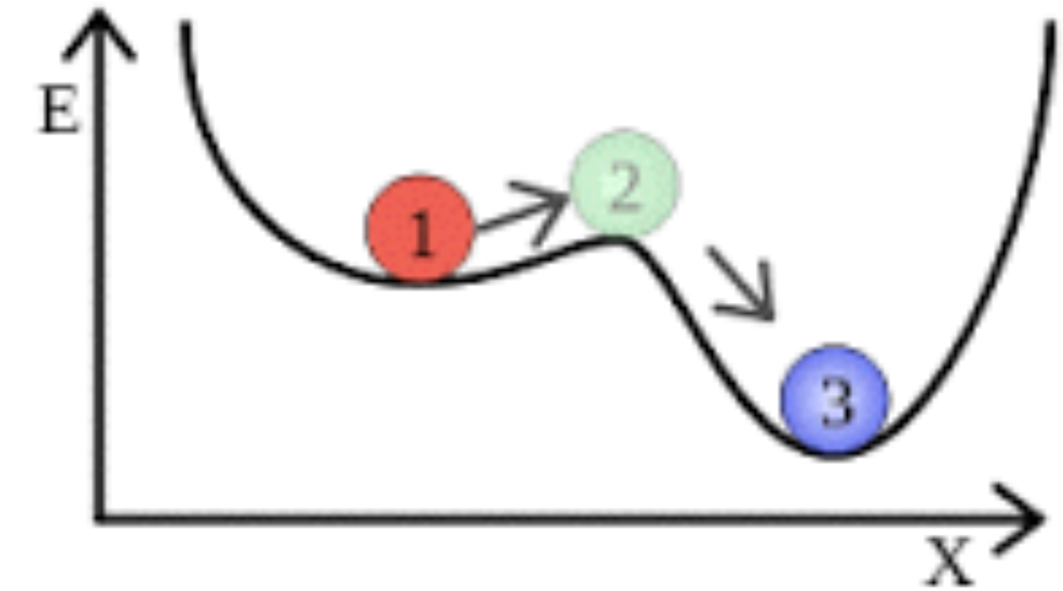
Pratik Chaudhari[1], Anna Choromanska[2], Stefano Soatto[1], Yann LeCun[2,3], Carlo Baldassi[4], Christian Borgs[5], Jennifer Chayes[5], Levent Sagun[2], Riccardo Zecchina[4]

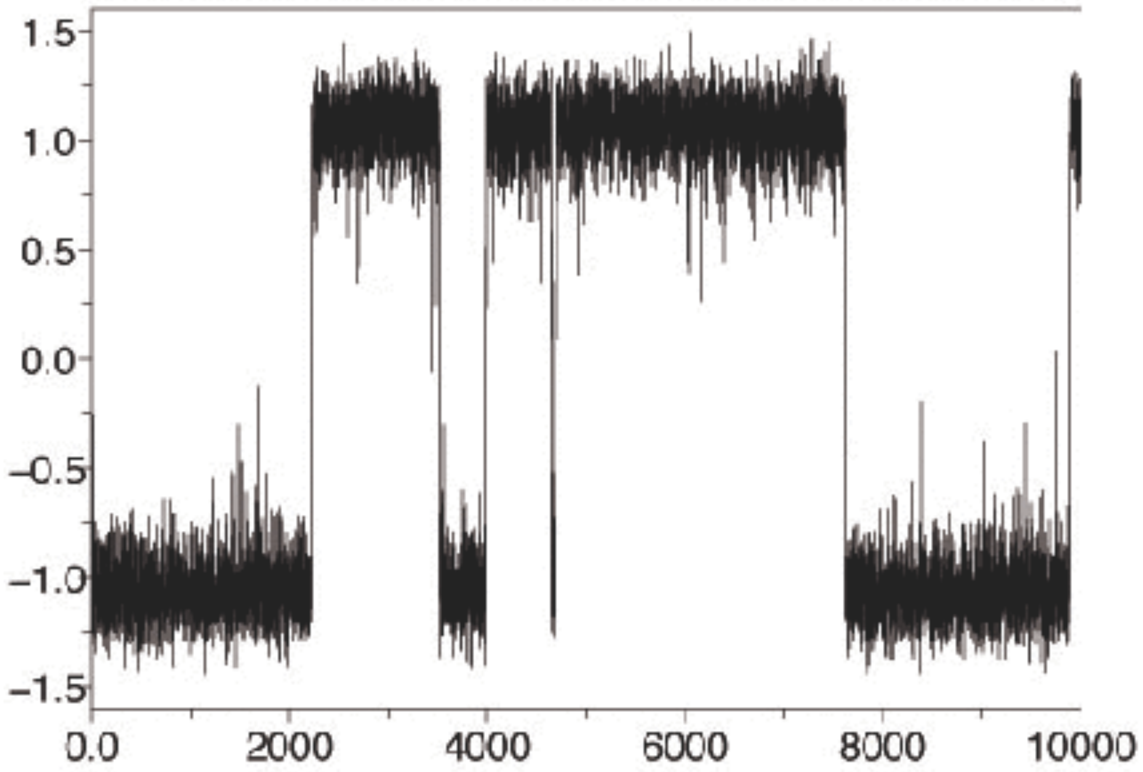
2017

Metastability:

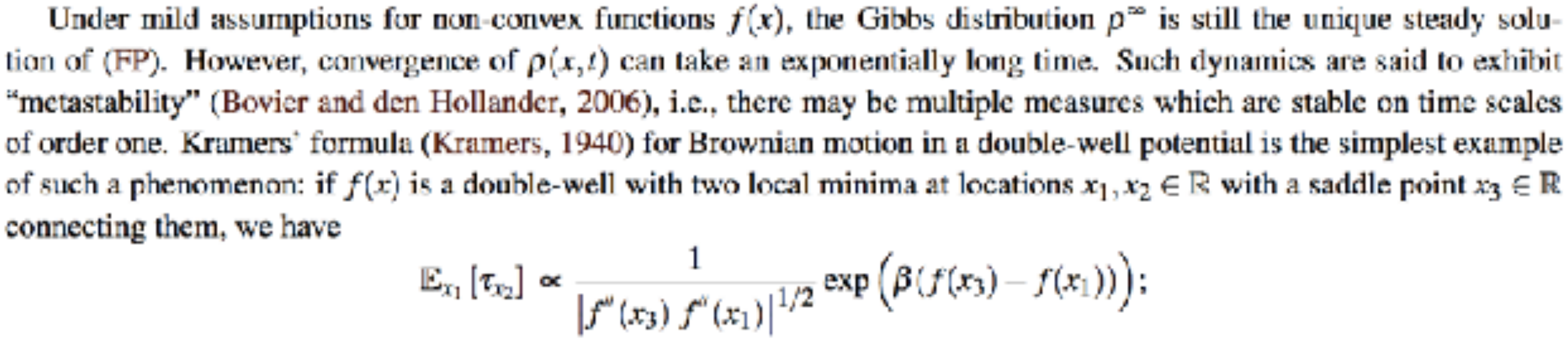
Exponential time to discover nearby minima

$$dx(t) = -\nabla f(x(t))\, dt + \sqrt{2\beta^{-1}}\, dW(t);$$





https://www.youtube.com/watch?v=Fot3m7kyLn4



Under mild assumptions for non-convex functions $f(x)$, the Gibbs distribution ρ^∞ is still the unique steady solution of (FP). However, convergence of $\rho(x,t)$ can take an exponentially long time. Such dynamics are said to exhibit "metastability" (Bovier and den Hollander, 2006), i.e., there may be multiple measures which are stable on time scales of order one. Kramers' formula (Kramers, 1940) for Brownian motion in a double-well potential is the simplest example of such a phenomenon: if $f(x)$ is a double-well with two local minima at locations $x_1, x_2 \in \mathbb{R}$ with a saddle point $x_3 \in \mathbb{R}$ connecting them, we have

$$\mathbb{E}_{x_1}[\tau_{x_2}] \propto \frac{1}{|f''(x_3)\, f''(x_1)|^{1/2}} \exp\Big(\beta\big(f(x_3) - f(x_1)\big)\Big);$$

For time O(1), there appear to be two invariant measures. Really there is one, but it is metastable.

Conclusion: SGD for nonconvex is challenging

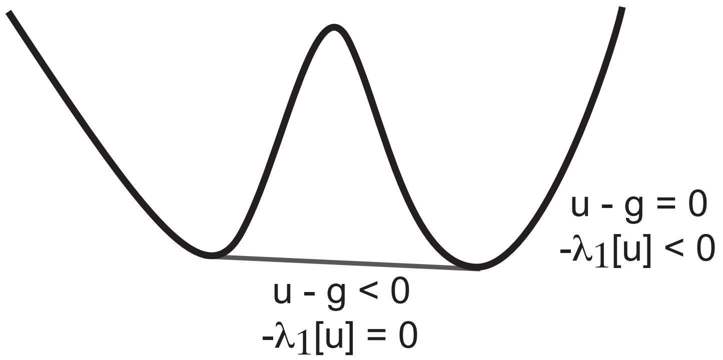
Metastability:
Discussion

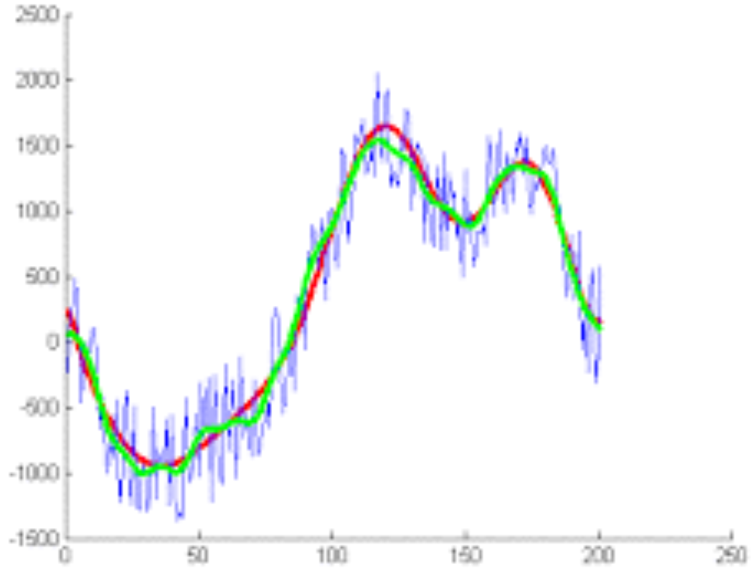
$$dx(t) = -\nabla f(x(t))\, dt + \sqrt{2\beta^{-1}}\, dW(t);$$

- Small steps: you fill find only local minima nearby
- Large steps: you may land in a better place, but how do be sure to stay nearby?
- Insight from Energy Landscape of f.
- Mathematics Problem: how do you learn about the geometry of f
- Algorithm: how to modify f so that it has a better landscape?

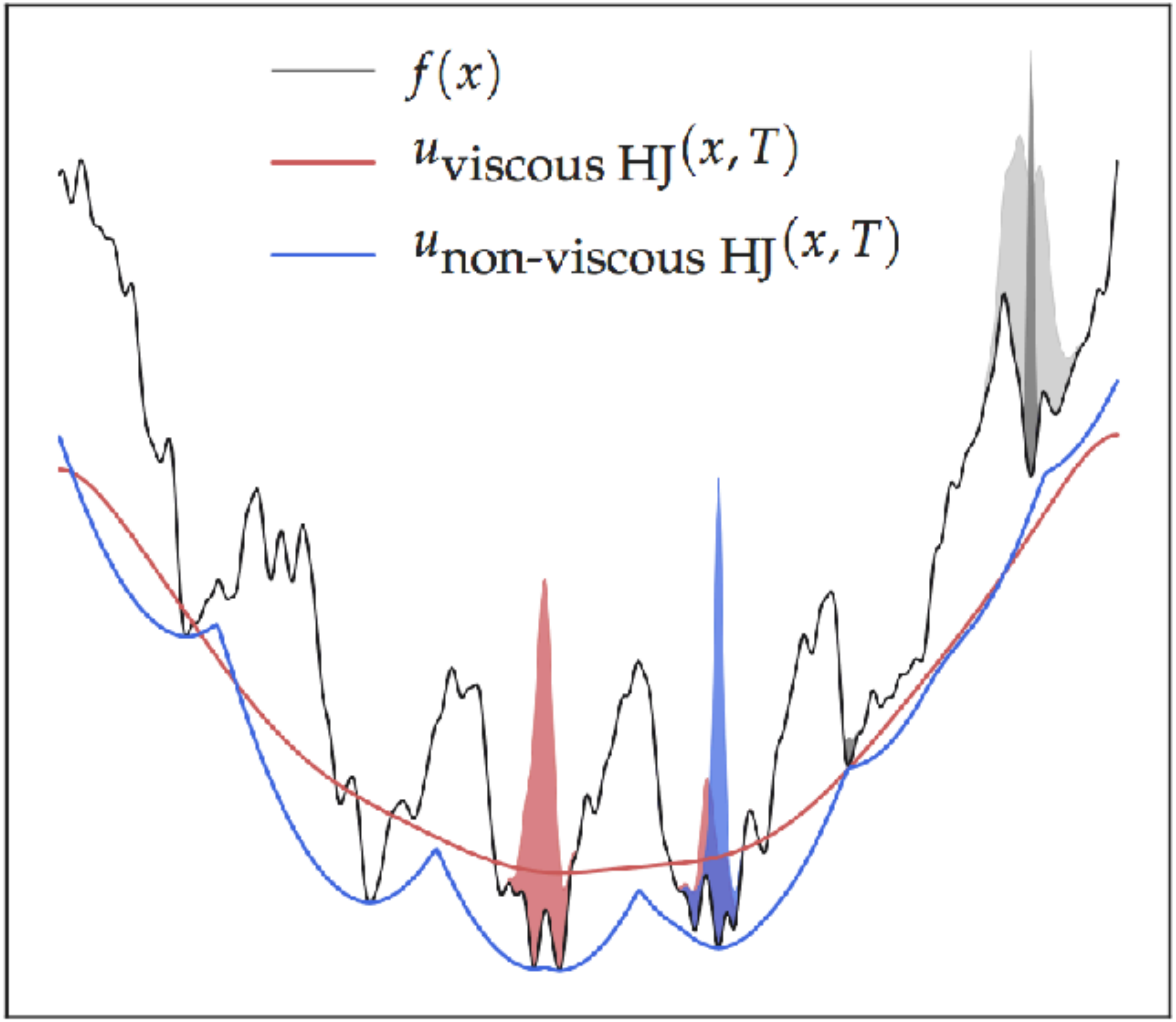
Idea: Continuation Methods

- *Idea*: replace minimization problem with a different problem, with same (or better) minima:
  - Replace f(x) with convex envelope
  - “Scoping”: start with a smoothed version of f, and decrease smoothing as minimization progresses.
- *Problem*: not practical in high dimensions.





Regularization using Viscous Hamilton-Jacobi PDE



- True solution in one dimension. (Cartoon in high dimensions, because algorithm only works for shorter times.)

Spin Glasses
(statistical physics motivation)*

**incomprehensible*

Ising model

$$-H(\sigma) = \sum_{(i,j)} J \, \sigma_i \, \sigma_j$$

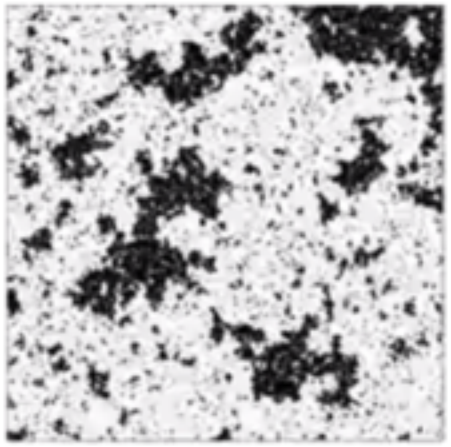
$\sigma_i \in \{-1, 1\}$

Hamiltonian

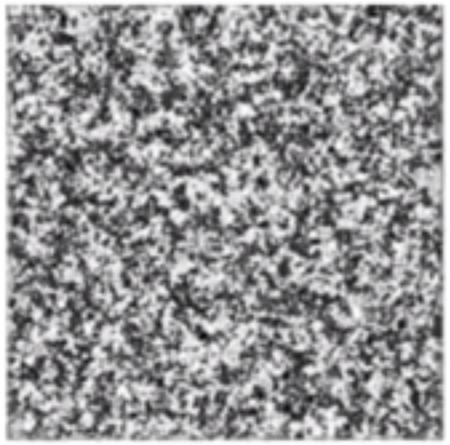
sum over all neighbors

coupling strength, also called "disorder"

upward / downward spins



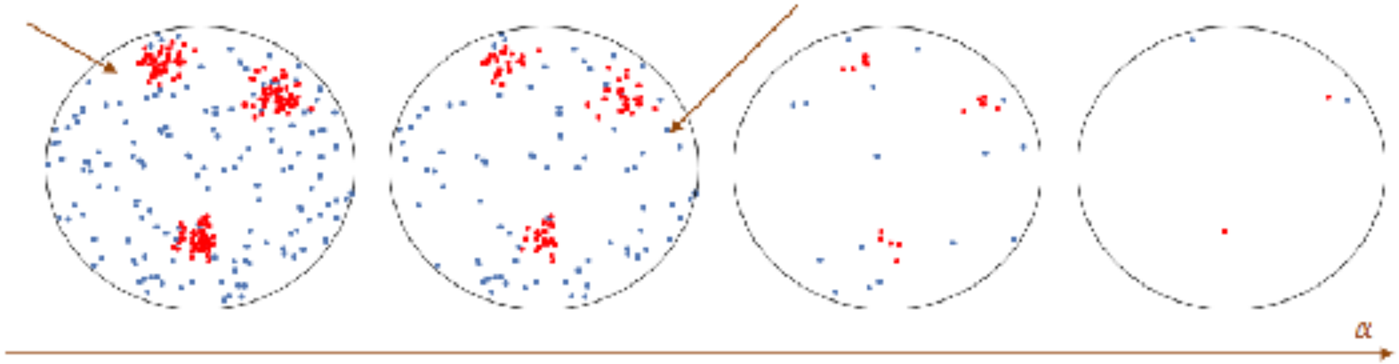
large correlations at
low temperature

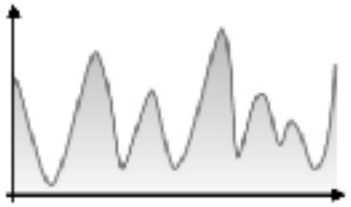


complete disorder at
high temperature

dense clusters do not show up in the
standard replica analysis of the
Gibbs distribution

only blue points = golf course-like landscape





"Local Entropy" improves the
energy landscape of non-convex
function.

Local entropy as a measure for sampling solutions in Constraint Satisfaction Problems

Carlo Baldassi,[1,2] Alessandro Ingrosso,[1,2] Carlo Lucibello,[1,2] Luca Saglietti,[1,2] and Riccardo Zecchina[1,2,3]

2016

What is Local Entropy?

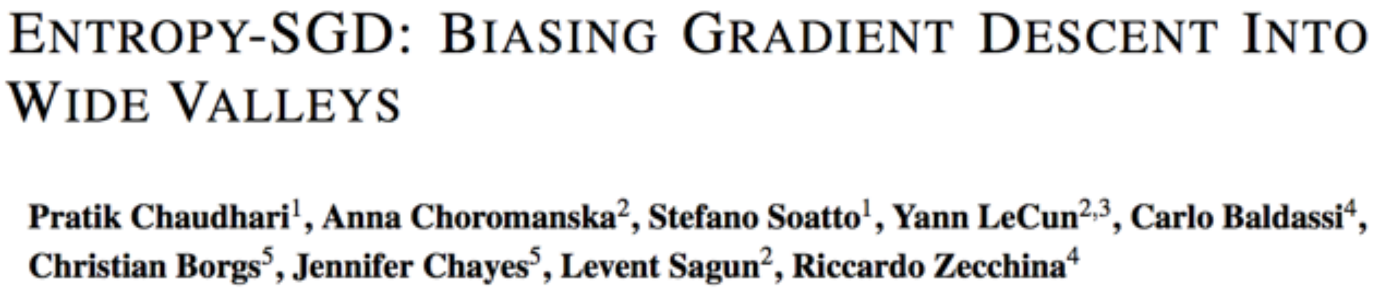
- Replace f(x) with "local entropy" of f

$$f_\gamma(x) = -\log\left[G_\gamma * e^{-f(x)}\right]$$

- *Algorithmic*: for small *gamma*, can evaluate grad f efficiently.

$$G_\gamma(x) = Ce^{-\frac{\|x\|^2}{2\gamma}},$$

- *Analytic:* can prove that scoping with "local entropy" improves the expected loss



ENTROPY-SGD: BIASING GRADIENT DESCENT INTO WIDE VALLEYS

Pratik Chaudhari[1], Anna Choromanska[2], Stefano Soatto[1], Yann LeCun[2,3], Carlo Baldassi[4], Christian Borgs[5], Jennifer Chayes[5], Levent Sagun[2], Riccardo Zecchina[4]

Jan 2017

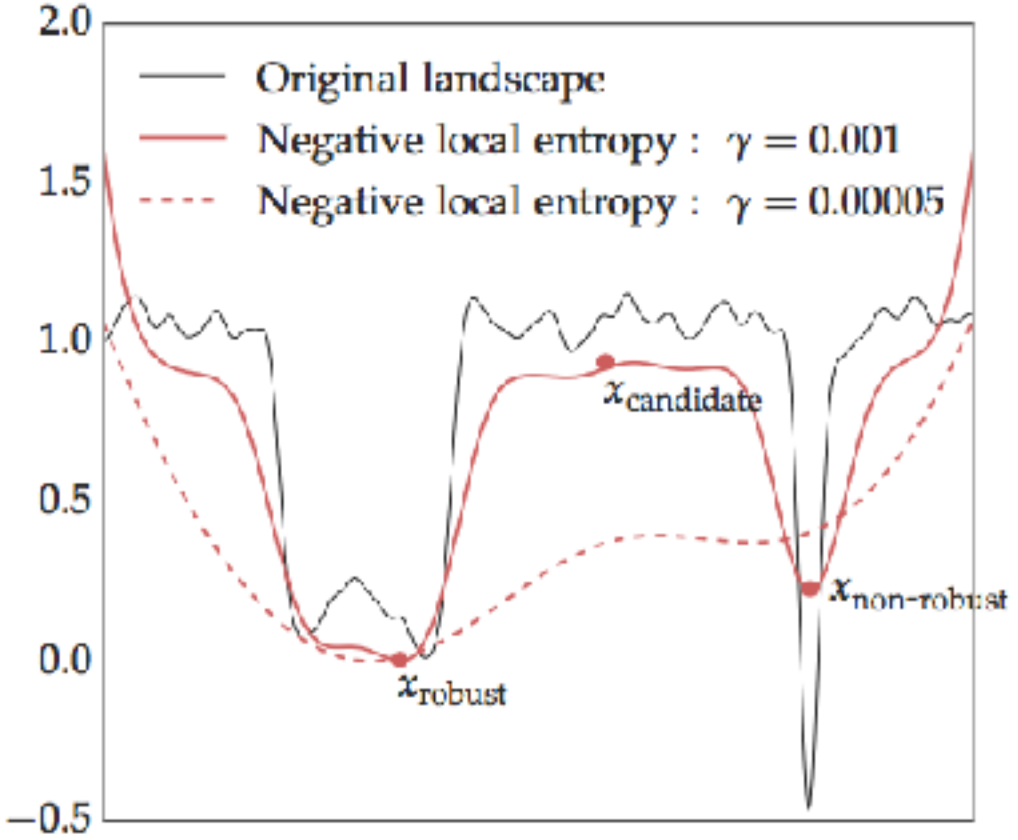


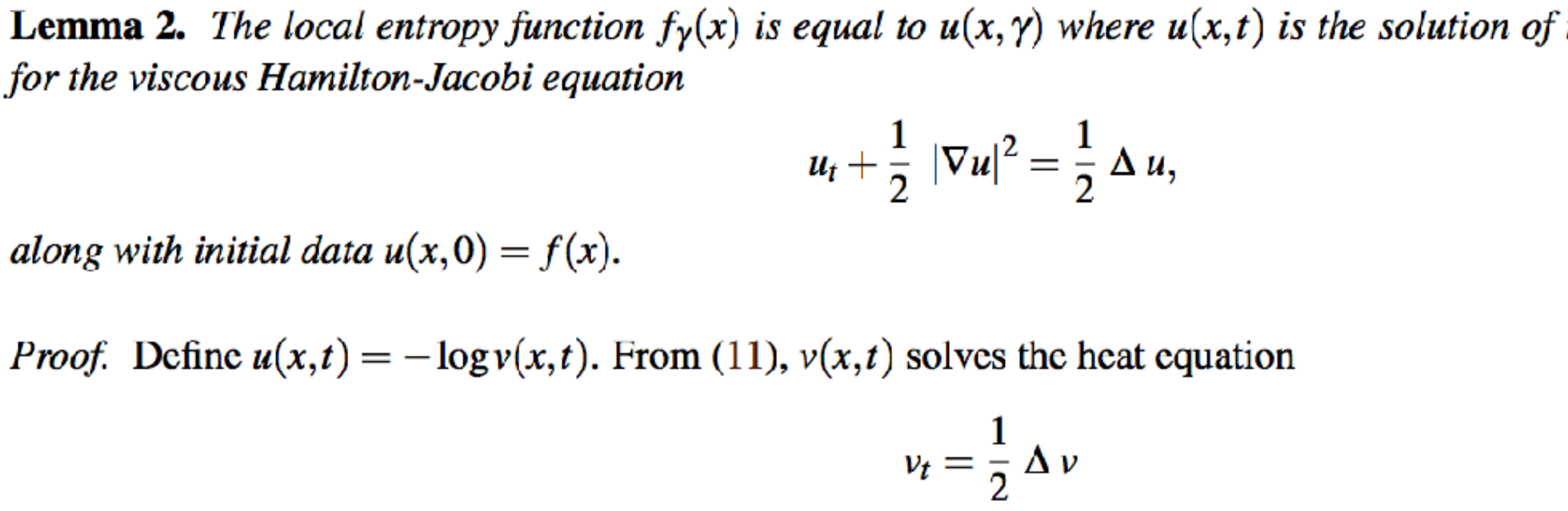
Figure 2: Local entropy concentrates on wide valleys in the energy landscape.

Outline

- 1. Introduction and Background
- 2. Math theory
    - PDE interpretation and derivation of Local Entropy
    - Proof of improvement
- 3. Algorithm and efficiency
- 4. Computational Results

(*Comment)* this is “Applied Math”: no new math theory. new *application* of existing math theory to problem. (We still prove theorems, but they are surprising to experts)

PDE for local entropy function and its gradient



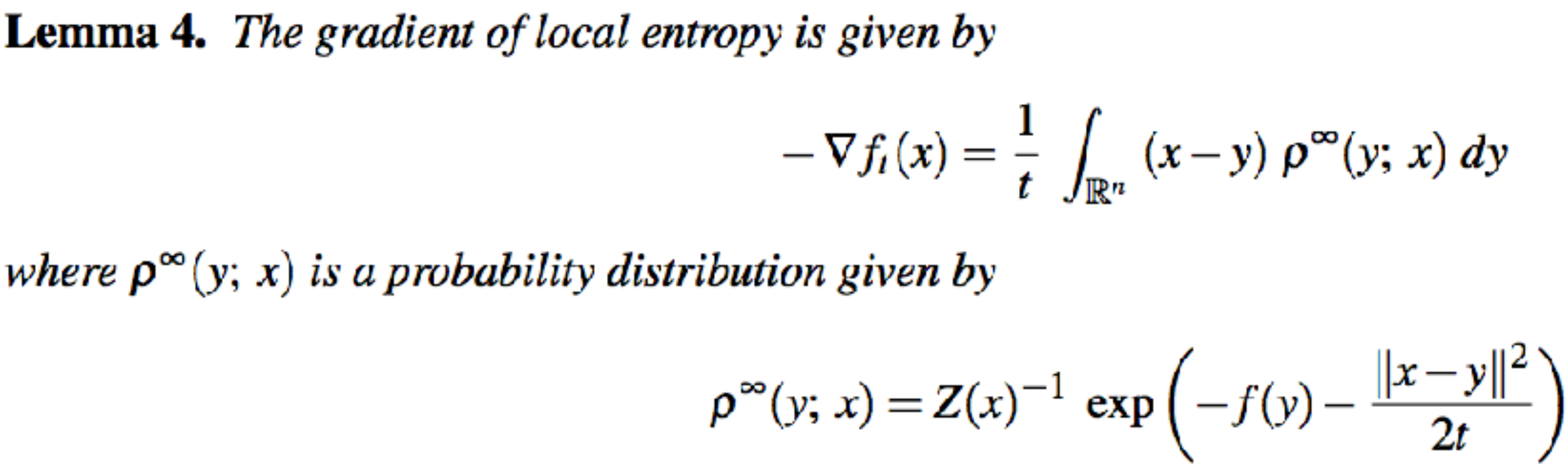
Lemma 2. *The local entropy function $f_\gamma(x)$ is equal to $u(x,\gamma)$ where $u(x,t)$ is the solution of for the viscous Hamilton-Jacobi equation*

$$u_t + \frac{1}{2}\ |\nabla u|^2 = \frac{1}{2}\ \Delta\, u,$$

along with initial data $u(x,0) = f(x)$.

Proof. Define $u(x,t) = -\log v(x,t)$. From (11), $v(x,t)$ solves the heat equation

$$v_t = \frac{1}{2}\ \Delta\, v$$



Lemma 4. *The gradient of local entropy is given by*

$$-\nabla f_t(x) = \frac{1}{t}\ \int_{\mathbb{R}^n} (x-y)\ \rho^\infty(y;\ x)\ dy$$

where $\rho^\infty(y;\ x)$ is a probability distribution given by

$$\rho^\infty(y;\ x) = Z(x)^{-1}\ \exp\left(-f(y) - \frac{\|x-y\|^2}{2t}\right)$$

Background: Stochastic Differential Equations

$$dx(t) = -\nabla f(x(t))\, dt + \sqrt{2\beta^{-1}}\, dW(t);$$

$$\mathcal{L}\phi = -\nabla f \cdot \nabla \phi + \beta^{-1}\Delta\, \phi. \qquad \mathcal{L}^*\rho = \nabla\cdot(\nabla f \rho) + \beta^{-1}\Delta\, \rho$$

$$\frac{\partial u}{\partial t} = \mathcal{L}u, \qquad \frac{\partial}{\partial t}\rho(x,t) = \mathcal{L}^*\rho$$

$$u(x,t) = \mathbb{E}\left[V(x(T)) \,\middle|\, x(t) = x\right] \qquad \rho^\infty(x;\, \beta) = Z(\beta)^{-1} e^{-\beta f(x)}$$

Background: JKO Functional and convergence rate in convex case

$$J(\rho) = \int f(x)\, \rho\, dx + \beta^{-1} \int \rho\, \log \rho\, dx;$$

$$d_{W_2}(\rho(x,t),\, \rho^\infty) \le d_{W_2}(\rho(x,0),\, \rho^\infty)\, e^{-\lambda t}. \qquad \nabla^2 f(x) \ge \lambda,$$

Background: Homogenization of SDEs

Pavliotis and Stuart (2008, Chap. 10, 17)

- Two scale dynamics

$$dx(s) = h(x,\ y)\ ds$$
$$dy(s) = \frac{1}{\varepsilon}\, g(x,\ y)\ ds + \frac{1}{\sqrt{\varepsilon}}\ dW(s);$$

- Assume a unique invariant solution of the fast dynamics

$$\mathscr{L}_0^*\ \rho^\infty(y;\ x) = 0;$$

- In the limit, obtain homogenized dynamics

$$d\bar{x}(s) = \bar{h}(x)\ ds$$

- given by averaging against the invariant measure

$$\bar{h}(x) = \int h(x,y)\ \rho^\infty(dy;\ x).$$

- Ergodicity:

$$= \lim_{T\to\infty} \frac{1}{T} \int_0^T h(x,y(s))\ ds$$

Derivation of Local Entropy using Homog SDEs

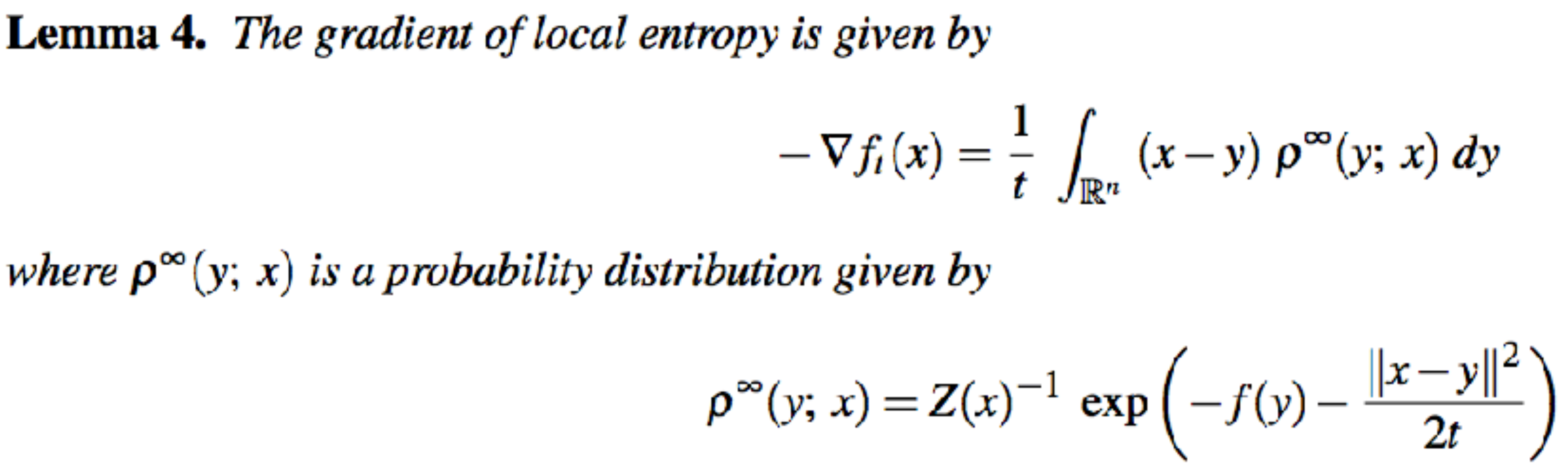
(ESGD)
Entropy SGD

$$dx(s) = -\gamma^{-1}\,(x-y)\,ds$$
$$dy(s) = -\frac{1}{\varepsilon}\left[\nabla f(y) + \frac{1}{\gamma}\,(y-x)\right]\,ds + \frac{1}{\sqrt{\varepsilon}}\,dW(s).$$

- Solve the FP equation, to obtain formula for invariant measure

$$H(x,y;\,\gamma) = f(y) + \tfrac{1}{2\gamma}\|x-y\|^2. \qquad \rho^\infty(y;\,x) \propto \exp\left(-H(x,y;\,\gamma)/\varepsilon\right)$$

- Apply the homogenization result (and soln of FP) to see ESGD gives the gradient



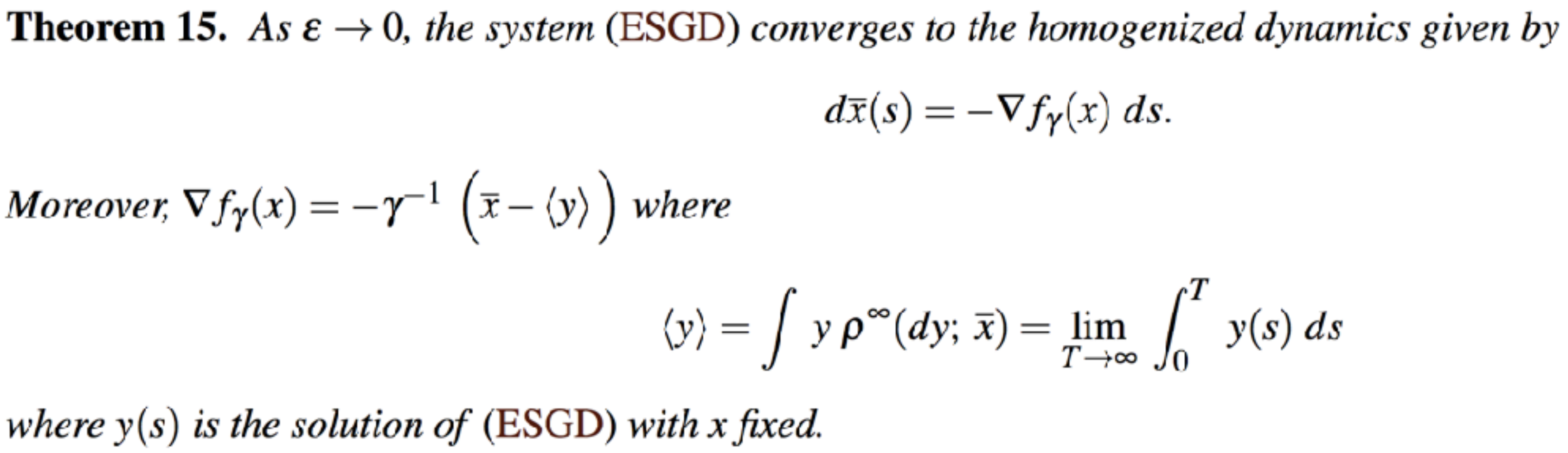
Lemma 4. *The gradient of local entropy is given by*

$$-\nabla f_\iota(x) = \frac{1}{t}\int_{\mathbb{R}^n}(x-y)\,\rho^\infty(y;\,x)\,dy$$

where $\rho^\infty(y;\,x)$ *is a probability distribution given by*

$$\rho^\infty(y;\,x) = Z(x)^{-1}\,\exp\left(-f(y) - \frac{\|x-y\|^2}{2t}\right)$$

Algorithmic Implication



Theorem 15. *As $\varepsilon \to 0$, the system* (ESGD) *converges to the homogenized dynamics given by*

$$d\bar{x}(s) = -\nabla f_\gamma(x)\, ds.$$

Moreover, $\nabla f_\gamma(x) = -\gamma^{-1}\left(\bar{x} - \langle y \rangle\right)$ *where*

$$\langle y \rangle = \int y\, \rho^\infty(dy;\, \bar{x}) = \lim_{T\to\infty} \int_0^T y(s)\, ds$$

where $y(s)$ is the solution of (ESGD) *with x fixed.*

- We showed that we can compute the gradient of local entropy function using an auxiliary SDE, provided the parameter gamma is chosen so that the function H = f + quadratic is convex

- The fast dynamics correspond to an “inner loop” of the algorithm. In practise, can take a handful of steps. Then the update is given by averaging the values of y from the inner loop

Outline

- 1. Introduction and Background
- 2. Math theory
  - PDE interpretation and derivation of Local Entropy
  - Proof of improvement
- 3. Algorithm and efficiency
- 4. Computational Results

Context for Improvement

- In the derivation, gamma was fixed. Better to "scope", sending gamma to 0 as outer loop progresses.

- We prove that scoping improves the expected value of a loss function, and quantify the improvement using dynamic programming

- This corresponds to nonlinear forward -backward equations

- Familiar in Mean-Field-Games

Forward-backward equations.

$$\frac{\partial u}{\partial t} = -\frac{1}{2}|\nabla u|^2 + \frac{1}{2}\Delta u$$

$$\rho_t = -\nabla \cdot \left(\nabla u \, \rho\right) + \Delta \rho,$$

$$u(x,T) = V(x),$$

$$\rho(x,0) = \rho_0(x).$$

Visualization of Improvement: dimension 1, PDE simulation.

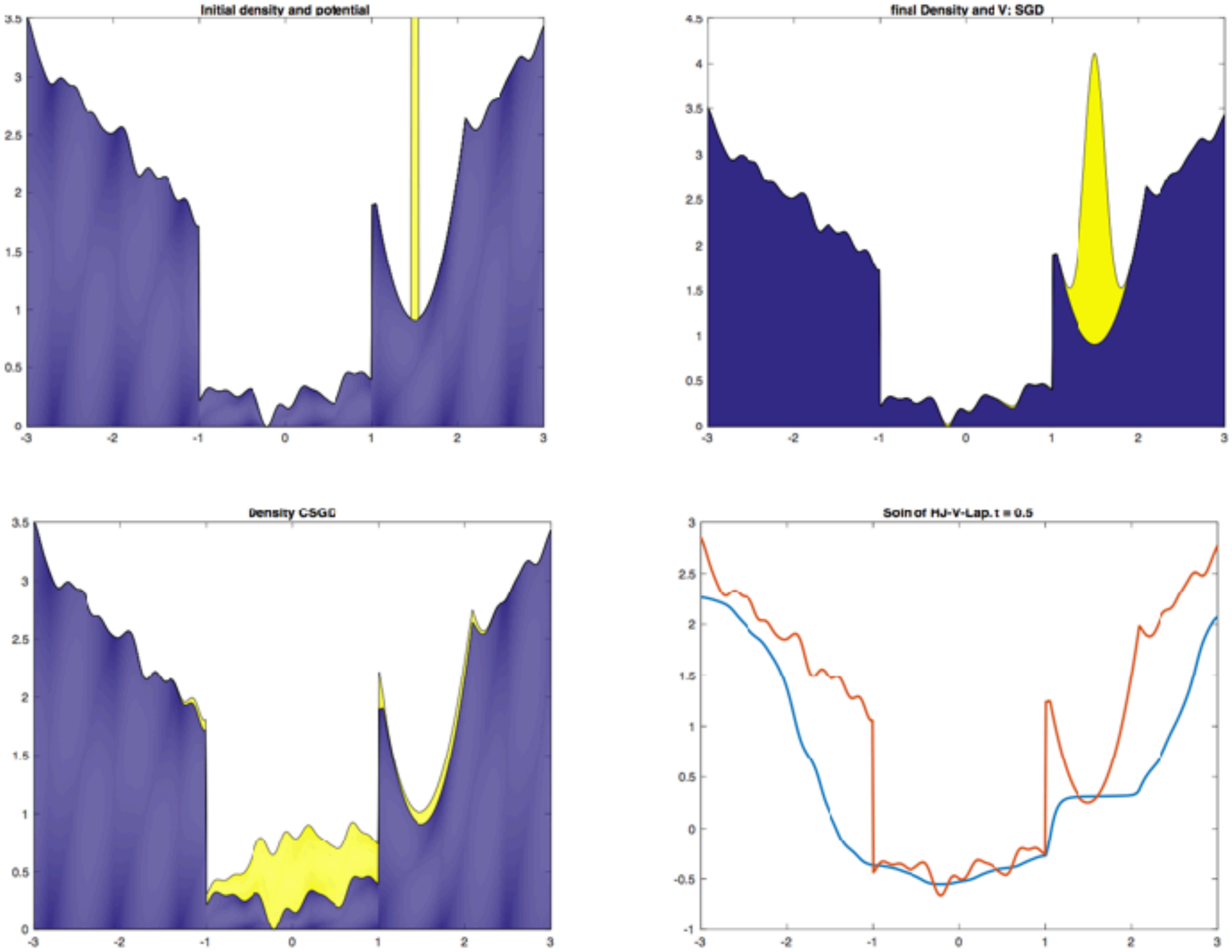
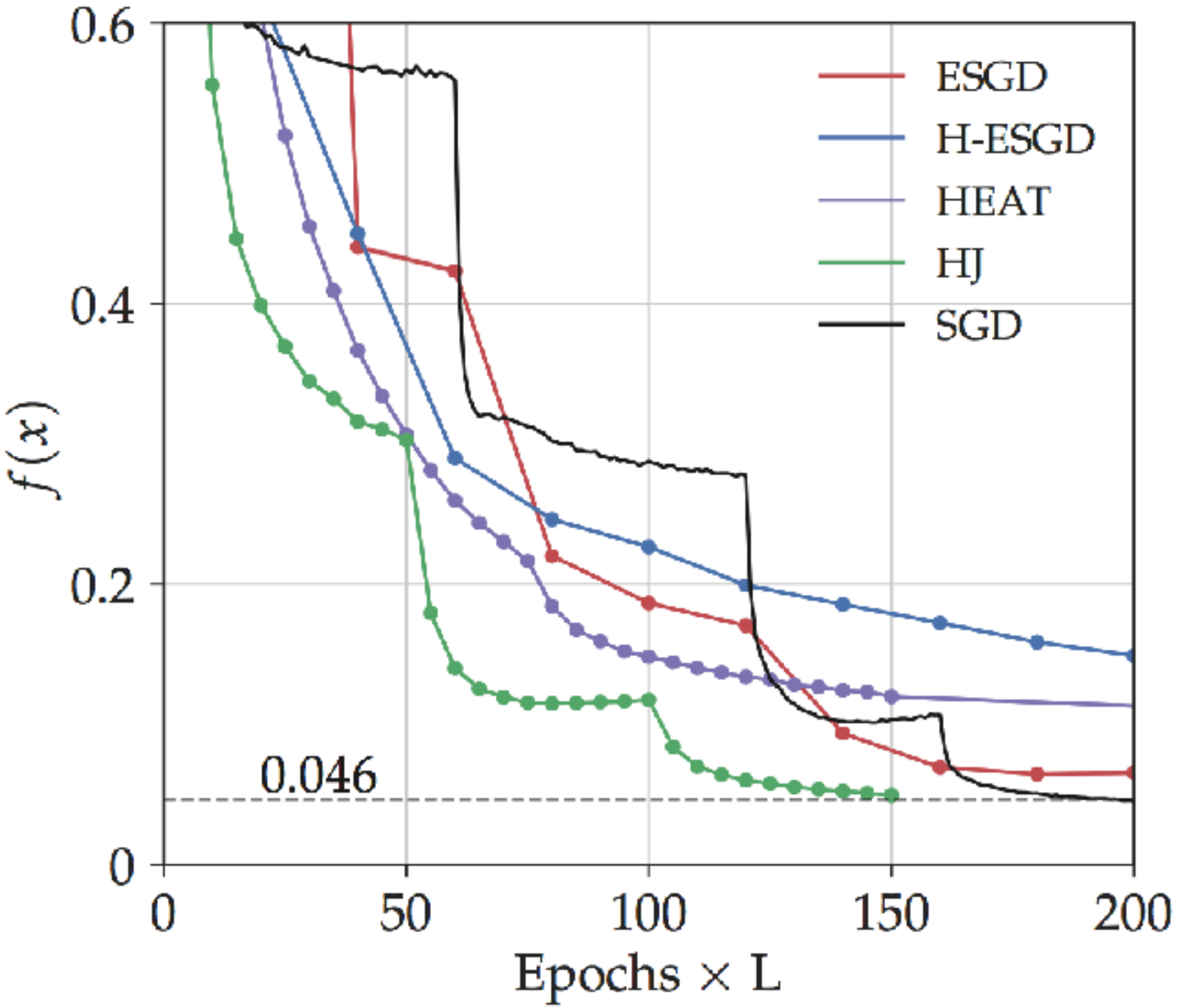


FIGURE 1. Initial density, final density SGD, final density CSGD, Solution of HJB **Forward-backward equations.**

Visualization of Improvement: single simulation dimension = 1.67 million



(A) All-CNN: Training loss

Review of controlled SDEs

Fleming and Rishel, 2012)

- SDE with control parameter

$$dx(s) = -h(x)\,ds - c_\alpha\,\alpha(s)\,ds + \sqrt{2\beta^{-1}}\,dW(s)$$
$$x(0) = x.$$

- Generator depends on the parameter

$$\mathscr{L}_\alpha := -(h + c_\alpha\,\alpha)\cdot\nabla + \beta^{-1}\,\Delta$$

- introduce cost functional

$$\mathscr{C}(x(\cdot),\,\alpha(\cdot)) = \mathbb{E}\left[V(x(T)) + \frac{c_\alpha}{2}\int_0^T \|\alpha(s)\|^2\,ds\right]$$

- define value function

$$u(x,t) = \min_{\alpha(\cdot)}\,\mathscr{C}(x(\cdot),\,\alpha(\cdot)).$$

- Value fn satisfies HJB

$$-u_t = -h\cdot\nabla u - \frac{c_\alpha}{2}\,|\nabla u|^2 + \beta^{-1}\Delta\,u,$$
$$u(x,T) = V(x).$$

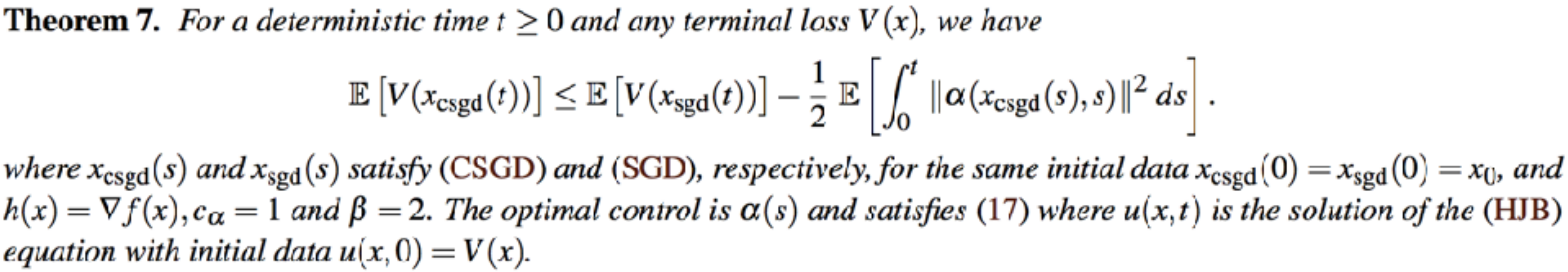
- Use DPP to select optimal control

$$\alpha(x,t) = \nabla u(x,t).$$

Proof of Improvement

- The proof of improvement uses stochastic optimal control theory.
- This is a technique to interpret local entropy as the value function of a stochastic optimal control problem.
- Allows us to conclude that we improve the expected value of f

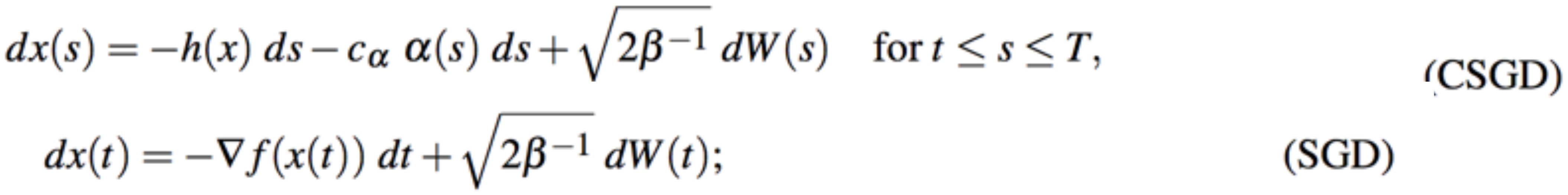
Expected Improvement Theorem



Theorem 7. *For a deterministic time $t \geq 0$ and any terminal loss $V(x)$, we have*

$$\mathbb{E}\left[V(x_{\mathrm{csgd}}(t))\right] \leq \mathbb{E}\left[V(x_{\mathrm{sgd}}(t))\right] - \frac{1}{2}\,\mathbb{E}\left[\int_0^t \|\alpha(x_{\mathrm{csgd}}(s),s)\|^2\,ds\right].$$

where $x_{\mathrm{csgd}}(s)$ and $x_{\mathrm{sgd}}(s)$ satisfy (CSGD) and (SGD), respectively, for the same initial data $x_{\mathrm{csgd}}(0) = x_{\mathrm{sgd}}(0) = x_0$, and $h(x) = \nabla f(x)$, $c_\alpha = 1$ and $\beta = 2$. The optimal control is $\alpha(s)$ and satisfies (17) where $u(x,t)$ is the solution of the (HJB) equation with initial data $u(x,0) = V(x)$.



$$dx(s) = -h(x)\,ds - c_\alpha\,\alpha(s)\,ds + \sqrt{2\beta^{-1}}\,dW(s) \quad \text{for } t \leq s \leq T, \tag{CSGD}$$

$$dx(t) = -\nabla f(x(t))\,dt + \sqrt{2\beta^{-1}}\,dW(t); \tag{SGD}$$

- Compares the expected value of SGD, with that of optimally controlled SGD.
- No surprise that adding a control improves the value function.
- The point is that we can interpret the scoping for local entropy as solving the same HJB equation, so we obtain a proof of improvement.
- The control is equal to the gradient.

Outline

- 1. Introduction and Background
- 2. Math theory
- 3. Algorithm and efficiency
- 4. Computational Results

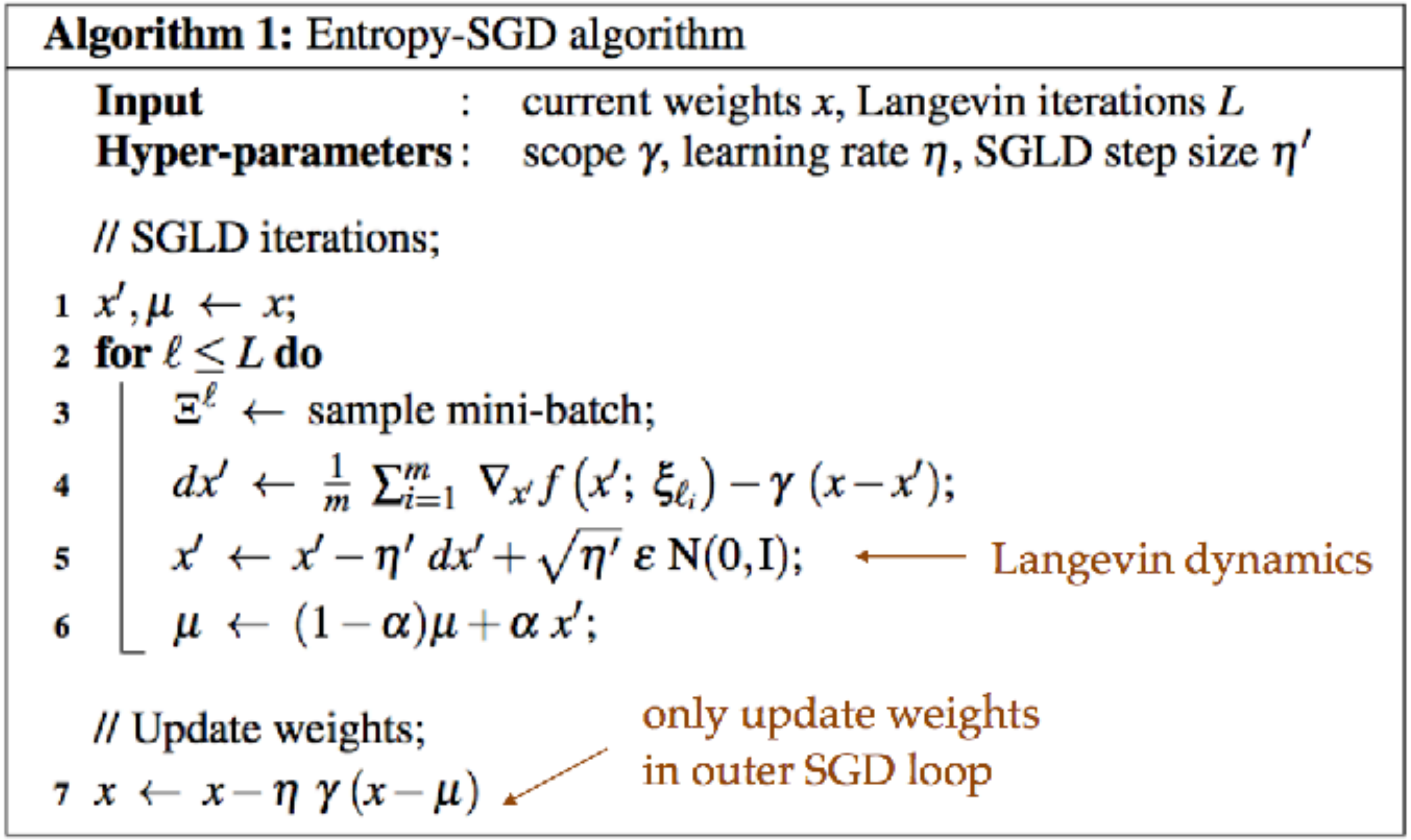
Algorithm for Local Entropy*

**will be translated into math*

- **Two nested for-loops of SGD**
  - outer loop updates parameters
  - inner loop estimates gradient of local entropy
- **Scoping: anneal the scoping parameter γ to search for minima**

Extremely effective —

1. searches parameter space in the beginning of training
2. Focusses on dense sub-clusters towards the end



Algorithm 1: Entropy-SGD algorithm

```
Input            : current weights $x$, Langevin iterations $L$
Hyper-parameters : scope $\gamma$, learning rate $\eta$, SGLD step size $\eta'$

// SGLD iterations;
$x', \mu \leftarrow x$;
for $\ell \leq L$ do
    $\Xi^{\ell} \leftarrow$ sample mini-batch;
    $dx' \leftarrow \frac{1}{m} \sum_{i=1}^{m} \nabla_{x'} f\left(x';\, \xi_{\ell_i}\right) - \gamma\,(x - x')$;
    $x' \leftarrow x' - \eta'\, dx' + \sqrt{\eta'}\, \varepsilon\, \mathrm{N}(0, \mathrm{I})$;
    $\mu \leftarrow (1-\alpha)\mu + \alpha\, x'$;

// Update weights;
$x \leftarrow x - \eta\, \gamma\, (x - \mu)$
```

Langevin dynamics

only update weights in outer SGD loop

Math Translation

- Scoping: this is the stochastic control interpretation, forward and backward equations.

- Inner and Outer Loop corresponds to Stochastic Homogenization of a 2-scale problem

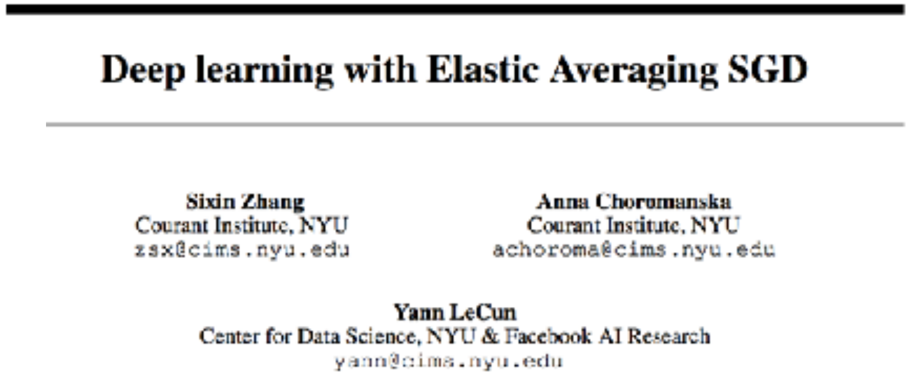
- Langevin dynamics - we actually have momentum, which is ignored for the purpose of the proof (and may not be needed in the algorithm)

- Comment: when gamma is not strong enough to make objective convex, algorithm does not compute local entropy. So physics intuition/PDE analysis is not consistent with algorithm. However, small quadratic term does not hurt.

Distributed Version of Algorithm

$$dx(s) = \frac{p^{-1}}{\gamma} \sum_{i=1}^{p} (x - y_i) \; ds$$

$$dy_i(s) = -\frac{1}{\varepsilon}\left[\nabla f(y_i) + \frac{1}{\gamma}\,(y_i - x)\right] ds + \frac{1}{\sqrt{\varepsilon}}\, dW_i(s); \quad \text{for } 1 \le i \le p.$$



Deep learning with Elastic Averaging SGD

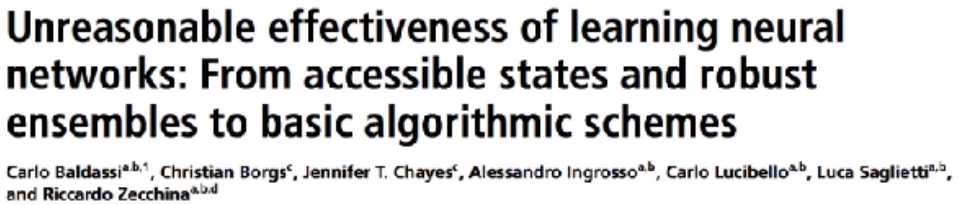
Sixin Zhang
Courant Institute, NYU
zsx@cims.nyu.edu

Anna Choromanska
Courant Institute, NYU
achoroma@cims.nyu.edu

Yann LeCun
Center for Data Science, NYU & Facebook AI Research
yann@cims.nyu.edu

- many workers on parallel processors each doing SGD, with an "elastic" coupling

- Similar argument, combining time average with spatial average, shows this algorithm is equivalent to computing the gradient of local entropy



Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes

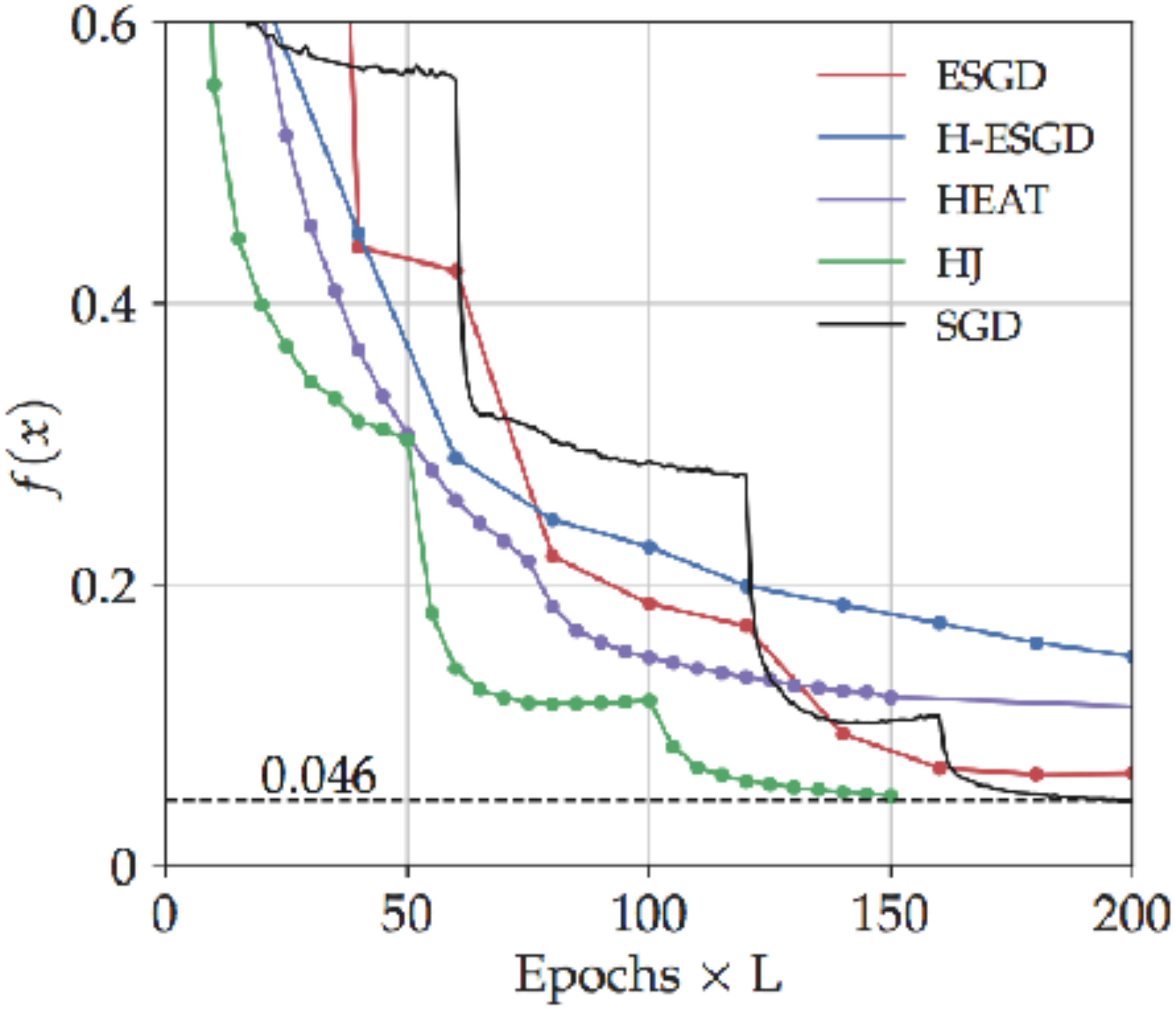
Carlo Baldassi[a,b,1], Christian Borgs[c], Jennifer T. Chayes[c], Alessandro Ingrosso[a,b], Carlo Lucibello[a,b], Luca Saglietti[a,b], and Riccardo Zecchina[a,b,d]

- The connection between these algorithms was only observed heuristically using stat physics Replica Theory. PDE/stochastic homogenization approach makes this very clear.

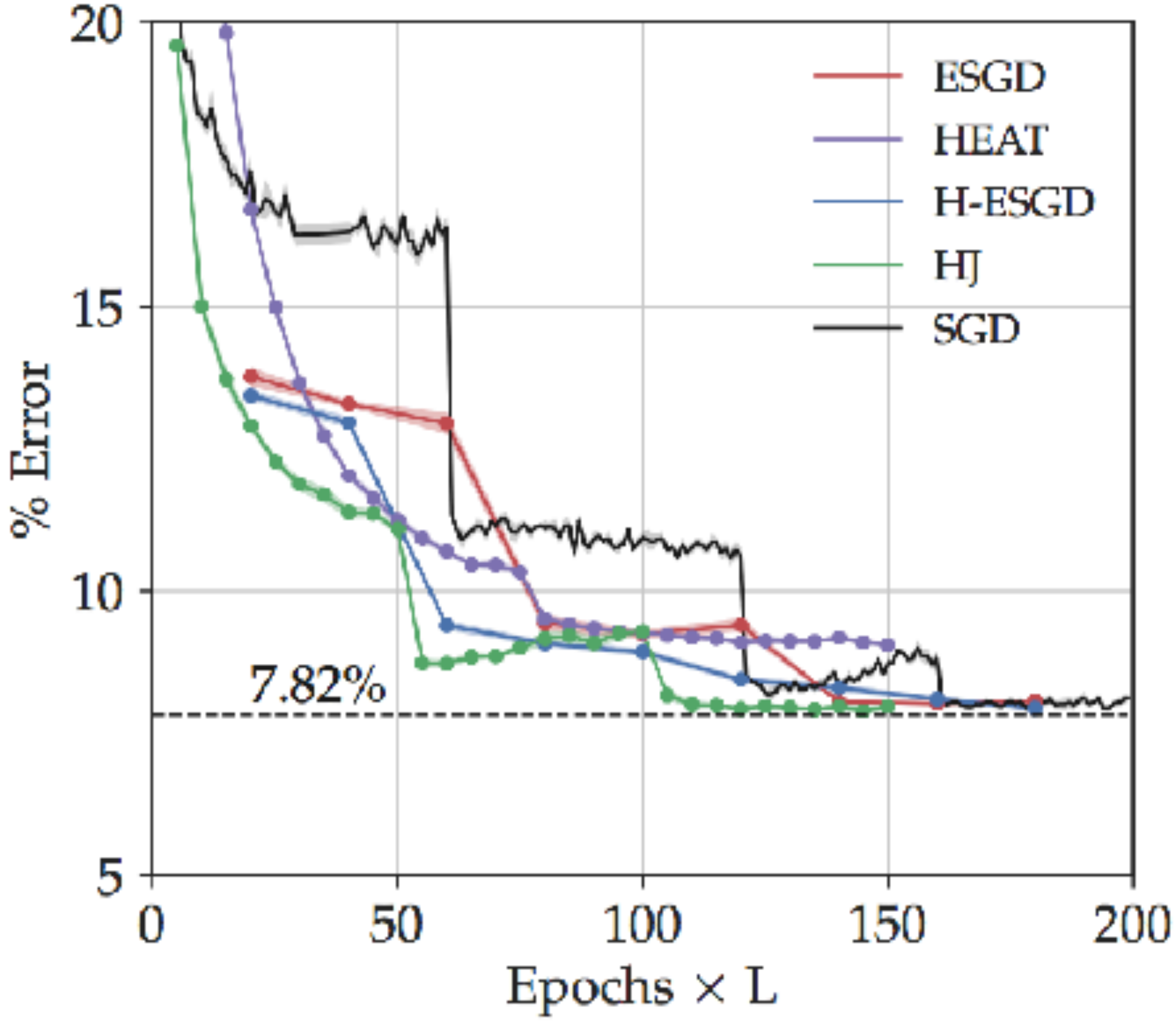
Outline

- 1. Introduction and Background
- 2. Math theory
- 3. Algorithm and efficiency
- 4. Computational Results

Numerical Results

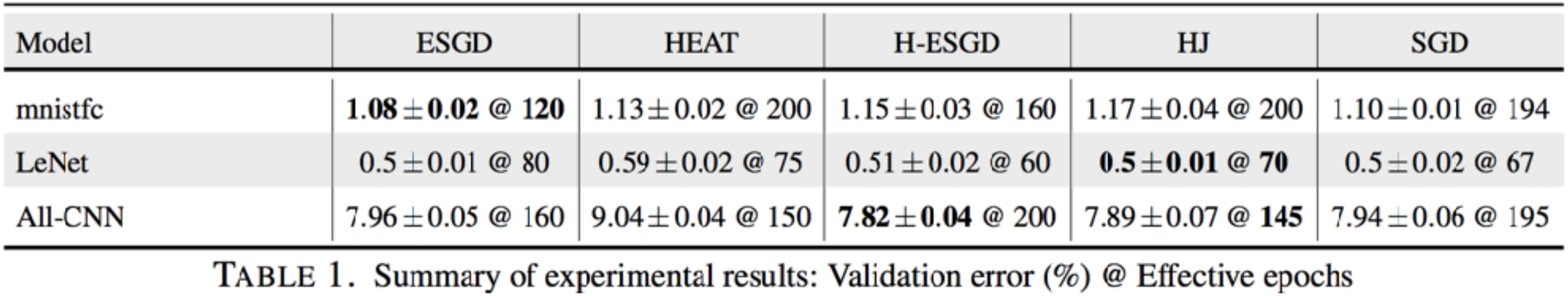


(A) All-CNN: Training loss



(B) All-CNN: Validation error

Numerical Results



Model	ESGD	HEAT	H-ESGD	HJ	SGD
mnistfc	**1.08 ± 0.02 @ 120**	1.13 ± 0.02 @ 200	1.15 ± 0.03 @ 160	1.17 ± 0.04 @ 200	1.10 ± 0.01 @ 194
LeNet	0.5 ± 0.01 @ 80	0.59 ± 0.02 @ 75	0.51 ± 0.02 @ 60	**0.5 ± 0.01 @ 70**	0.5 ± 0.02 @ 67
All-CNN	7.96 ± 0.05 @ 160	9.04 ± 0.04 @ 150	**7.82 ± 0.04** @ 200	7.89 ± 0.07 @ **145**	7.94 ± 0.06 @ 195

TABLE 1. Summary of experimental results: Validation error (%) @ Effective epochs