PDE and control methods for global optimization in deep neural networks

Martino Bardi [1] and Hicham Kouhkouh [2]

[1] Dipartimento di Matematica "Tullio Levi-Civita"
Università di Padova

[2] RTG Energy, Entropy, and Dissipative Dynamics
Institute for Mathematics, RWTH Aachen University

Nonlinear PDEs: theory, methods and applications

in memory of Maurizio Falcone

Università Roma "La Sapienza", May 24-26, 2023

Dynamic Games and Applications
Special Issue on:
Optimal control and differential games: theory, numerics and applications.
In memory of Maurizio Falcone.

Guest editors:
Martino Bardi, Università di Padova, Italy, bardi@math.unipd.it
Fabio Camilli, Università di Roma "La Sapienza", Italy, camilli@dmmm.uniroma1.it
Francisco José Silva Álvarez, Université de Limoges, France, francisco.silva@unilim.fr

Dynamic Games and Applications will publish a special issue on Optimal control and differential games: theory, numerics and applications, in memory of Maurizio Falcone (1954 - 2022) who made important contributions to these subjects.

The topics of the submitted articles could include approximation schemes for differential games, their convergence and numerical experiments, viscosity solutions of Hamilton-Jacobi-Bellman-Isaacs equations, as well as Mean-Field Games, but are not restricted to these subjects.

Submission Deadline: November 26th, 2023
Publication Date: Late 2024 – Early 2025

For submission instructions, please visit:
http://www.springer.com/mathematics/applications/journal/13235
Earlier submission is encouraged, and papers will appear online following acceptance in advance of the production of the full special issue.

Plan

- Global optimization
  - Entropic gradient descent
  - The Deep relaxation algorithm and Singular Perturbations
- Deep relaxation with control
  - Convergence of the value function
  - Convergence of trajectories.
- Methods:
  - Homogenization of the Hamilton-Jacobi-Bellman equation
  - The effective Hamiltonian via ergodic control
  - The limit is a value function

Global unconstrained optimization

Problem: Given a loss function $f \in C(\mathbb{R}^n)$, find a strategy (a dynamical system...) to reach the global minima of f (if they exist....).

Recent interest in this very classical problem comes from deep learning in neural networks: n very large, f highly nonlinear, non-convex and non-smooth....

Easy case: When f is convex and smooth, then the gradient flow or gradient descent (GD) answers the problem, i.e., any trajectory of

$$\dot{y}(s) = -\nabla f(y(s))$$

tends to argmin f as $t \to +\infty$.

In general a trajectory converges to a local minimum or a saddle point.

Global unconstrained optimization

Problem: Given a loss function $f \in C(\mathbb{R}^n)$, find a strategy (a dynamical system...) to reach the global minima of f (if they exist....).

Recent interest in this very classical problem comes from deep learning in neural networks: n very large, f highly nonlinear, non-convex and non-smooth....

Easy case: When f is convex and smooth, then the gradient flow or gradient descent (GD) answers the problem, i.e., any trajectory of

$$\dot{y}(s) = -\nabla f(y(s))$$

tends to $\operatorname{argmin} f$ as $t \to +\infty$.

In general a trajectory converges to a local minimum or a saddle point.

Classical variants for NON-convex functions f :

Stochastic gradient descent

$$dy(s) = -\nabla f(y(s))\, ds + \varepsilon dW_s,$$

W_s = Wiener process, avoids saddle points and shallow minima.
$-\nabla f$ = exploitation, εdW_s = exploration

Problems:

- no guarantee of convergence,
- ∇f may not exist... need to regularize f,
- a sufficiently low "robust" minimum, i.e., with large basin of attraction, can be preferable to a lower minimum in a narrow valley.

Classical variants for NON-convex functions f :

Stochastic gradient descent

$$dy(s) = -\nabla f(y(s))\, ds + \varepsilon dW_s,$$

$W_s =$ Wiener process, avoids saddle points and shallow minima.
$-\nabla f =$ exploitation, $\varepsilon dW_s =$ exploration

Problems:

- no guarantee of convergence,
- ∇f may not exist... need to regularize f,
- a sufficiently low "robust" minimum, i.e., with large basin of attraction, can be preferable to a lower minimum in a narrow valley.

Entropy regularization

Based on this understanding of how the local geometry looks at the end of optimization, can we modify SGD to actively seek such regions? Motivated by the work of Baldassi *et al* (2015) on shallow networks, instead of minimizing the original loss $f(x)$, we propose to maximize

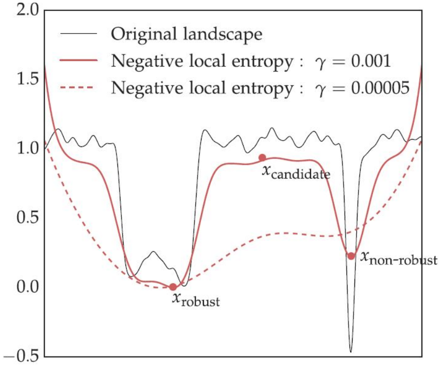
$$F(x, \gamma) = \log \int_{x' \in \mathbb{R}^n} \exp\left(-f(x') - \frac{\gamma}{2} \, ||x - x'||_2^2\right) \, dx'.$$

The above is a log-partition function that measures both the depth of a valley at a location $x \in \mathbb{R}^n$, and its flatness through the entropy of $f(x')$; we call it 'local entropy' in analogy to the free entropy used in statistical physics. The Entropy-SGD algorithm presented in this paper employs stochastic gradient Langevin dynamics (SGLD) to approximate the gradient of local entropy. Our algorithm resembles two nested loops of SGD: the inner loop consists of SGLD iterations while the outer loop updates the parameters. We show that the above modified loss function results in a smoother

Figure: From Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., ... & Zecchina, R. (2019). **Entropy-SGD: Biasing gradient descent into wide valleys.** J. Statistical Mechanics: Theory and Experiment, 2019(12), 124018.

From Chaudhari, LeCun et al J. Stat. Mech. (2019) :

“*We expect x_{robust} to be more robust that $x_{non\text{-}robust}$ to perturbations of data or parameters and thus generalize well (...). For low values of γ, the energy landscape is significantly smoother than the orginal landscape and still maintains our desired characteristic: global minimimum at a wide valley.*”



“*The local entropy thus provides a way of picking large, approximately flat, regions of the landscape over sharp, narrow valleys in spite of the latter possibly having a lower loss.*”

Local entropy and Gibbs distribution

From Chaudhari, Le Cun et al J. Stat. Mech. (2019):
"To focus on the flat regions such as x_{robust}, we construct ......

Definition (Local Entropy)

= the log-partition function of the modified Gibbs distribution:

$$F(x,\gamma) = \log \int_{\mathbb{R}^n} \exp\left(-f(y) - \frac{\gamma}{2}|x-y|^2\right) dy$$

$$= \log\left[\exp\left(-\frac{\gamma|\cdot|^2}{2}\right) * \exp(-f(x))\right].$$

$$\Rightarrow \qquad F(x,\gamma) = \log\Big(G_\gamma * \exp(-f(x))\Big),$$

where G_γ is the heat kernel (up to a multiplicative constant).

Local entropy and Gibbs distribution

From Chaudhari, Le Cun et al J. Stat. Mech. (2019):
"To focus on the flat regions such as x_{robust}, we construct ......

Definition (Local Entropy)

$=$ the log-partition function of the modified Gibbs distribution:

$$F(x,\gamma) = \log \int_{\mathbb{R}^n} \exp\left(-f(y) - \frac{\gamma}{2}|x-y|^2\right) \mathrm{d}y$$
$$= \log\left[\exp\left(-\frac{\gamma|\cdot|^2}{2}\right) * \exp(-f(x))\right].$$

$$\Rightarrow \qquad F(x,\gamma) = \log\Big(G_\gamma * \exp\left(-f(x)\right)\Big),$$

where G_γ is the heat kernel (up to a multiplicative constant).

Local entropy and Gibbs distribution

From Chaudhari, Le Cun et al J. Stat. Mech. (2019):
"To focus on the flat regions such as x_{robust}, we construct ......

Definition (Local Entropy)

= the log-partition function of the modified Gibbs distribution:

$$F(x,\gamma) = \log \int_{\mathbb{R}^n} \exp\left(-f(y) - \frac{\gamma}{2}|x-y|^2\right) dy$$
$$= \log\left[\exp\left(-\frac{\gamma|\cdot|^2}{2}\right) * \exp(-f(x))\right].$$

$$\Rightarrow \qquad F(x,\gamma) = \log\left(G_\gamma * \exp(-f(x))\right),$$

where G_γ is the heat kernel (up to a multiplicative constant).

The entropic gradient

A calculation gives the nice structure of the gradient:

$$\nabla_x F(x,\gamma) = \int_{\mathbf{R}^n} -\gamma(x-y)\,\rho_\infty(dy\,;x)$$

$$\rho_\infty(y;x) = \exp\left(-f(y) - \frac{\gamma}{2}|x-y|^2\right)/Z(x)$$

$(Z(x) =$ normalizing constant $) \Longrightarrow \nabla_x F(x,\gamma)$ is an **average** of $x-y$ w.r.t. the Gibbs measure ρ_∞ , that does not depend on ∇f.

Under mild assumptions on f and for γ large enough, the process

$$\text{(E)} \qquad dY_t = -\nabla_y\left(f(Y_t) + \frac{\gamma}{2}|x-Y_t|^2\right)dt + \sqrt{2}\,dW_t$$

is ergodic, and ρ_∞ is its invariant measure. Then for all initial positions Y_0 of (E)

$$\int_{\mathbf{R}^n} y\,\rho_\infty(dy\,;x) = \lim_{T\to+\infty}\frac{1}{T}\int_0^T \mathbb{E}\,[Y_t]\,dt.$$

The entropic gradient

A calculation gives the nice structure of the gradient:

$$\nabla_x F(x,\gamma) = \int_{\mathbf{R}^n} -\gamma(x-y)\,\rho_\infty(\mathrm{d}y\,;x)$$

$$\rho_\infty(y;x) = \exp\left(-f(y) - \frac{\gamma}{2}|x-y|^2\right)/Z(x)$$

($Z(x) =$ normalizing constant) $\Longrightarrow$ $\nabla_x F(x,\gamma)$ is an **average** of $x-y$ w.r.t. the Gibbs measure ρ_∞ , that does not depend on ∇f.

Under mild assumptions on f and for γ large enough, the process

$$\text{(E)} \qquad \mathrm{d}Y_t = -\nabla_y\left(f(Y_t) + \frac{\gamma}{2}|x-Y_t|^2\right)\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t$$

is ergodic, and ρ_∞ is its invariant measure. Then for all initial positions Y_0 of (E)

$$\int_{\mathbf{R}^n} y\,\rho_\infty(\mathrm{d}y\,;x) = \lim_{T\to+\infty}\frac{1}{T}\int_0^T \mathbb{E}\,[Y_t]\,dt.$$

Approximation of the entropic gradient descent

Problem: find an efficient approximation of

$$\dot{X}_t = \nabla_x F(X_t, \gamma) = \int_{\mathbf{R}^n} -\gamma(X_t - y)\, \rho_\infty(dy\,; X_t)$$

Difficulty: how to compute the average on the right hand side?

Chaudhari, P., Oberman, A., Osher, S., Soatto, S., Carlier, G. :

Deep relaxation: partial differential equations for optimizing deep neural networks. Res. Math. Sci. 2018,

propose an algorithm, called *Deep Relaxation* based on the system with different time scales

$$\begin{cases} dX_t^\varepsilon = -\gamma\,(X_t^\varepsilon - Y_t^\varepsilon)\, dt \\ dY_t^\varepsilon = -\frac{1}{\varepsilon}\nabla_y\Big(f(Y_t^\varepsilon) + \frac{\gamma}{2}|X_t^\varepsilon - Y_t^\varepsilon|^2\Big)dt + \sqrt{\frac{2}{\varepsilon}}\, dW_t \end{cases}$$

Y_t^ε = fast variables approximating $Y_.$ solving (E) for large times.

Approximation of the entropic gradient descent

Problem: find an efficient approximation of

$$\dot{X}_t = \nabla_x F(X_t, \gamma) = \int_{\mathbf{R}^n} -\gamma(X_t - y)\, \rho_\infty(dy\,; X_t)$$

Difficulty: how to compute the average on the right hand side?

Chaudhari, P., Oberman, A., Osher, S., Soatto, S., Carlier, G. :

Deep relaxation: partial differential equations for optimizing deep neural networks. Res. Math. Sci. 2018,

propose an algorithm, called *Deep Relaxation* based on the system with different time scales

$$\begin{cases} dX_t^\varepsilon = -\gamma\,(X_t^\varepsilon - Y_t^\varepsilon)\, dt \\ dY_t^\varepsilon = -\frac{1}{\varepsilon}\nabla_y\Big(f(Y_t^\varepsilon) + \frac{\gamma}{2}|X_t^\varepsilon - Y_t^\varepsilon|^2\Big)dt + \sqrt{\frac{2}{\varepsilon}}\, dW_t \end{cases}$$

$Y_t^\varepsilon =$ fast variables approximating $Y_\cdot$ solving (E) for large times.

The singular perturbation problem

Consider the coupled system with different time scales

$$\begin{cases} dX_t^\varepsilon = -\gamma\,(X_t^\varepsilon - Y_t^\varepsilon)\,dt \\ dY_t^\varepsilon = -\frac{1}{\varepsilon}\nabla_y\Big(f(Y_t^\varepsilon) + \frac{\gamma}{2}|X_t^\varepsilon - Y_t^\varepsilon|^2\Big)dt + \sqrt{\frac{2}{\varepsilon}}\,dW_t \end{cases}$$

where X_t^ε are "slow" variables and Y_t^ε "fast" variables.

Expect, for $\varepsilon \to 0$, $t/\varepsilon \to +\infty$,

$$Y_t^\varepsilon \approx Y_{t/\varepsilon} \approx \int_{\mathbf{R}^n} y\,\rho_\infty(dy\,;X_t^\varepsilon), \quad Y_\cdot \text{ solving (E)},$$

so by a homogenization procedure the two-scale system should converge, as $\varepsilon \to 0$, to the averaged system

$$\dot{X}_t = -\gamma\left(X_t - \int_{\mathbf{R}^n} y\,\rho_\infty(dy\,;X_t)\right) = \nabla_x F(X_t,\gamma)$$

which is the entropic gradient descent !

The singular perturbation problem

Consider the coupled system with different time scales

$$\begin{cases} dX_t^\varepsilon = -\gamma\,(X_t^\varepsilon - Y_t^\varepsilon)\,dt \\ dY_t^\varepsilon = -\frac{1}{\varepsilon}\nabla_y\Big(f(Y_t^\varepsilon) + \frac{\gamma}{2}|X_t^\varepsilon - Y_t^\varepsilon|^2\Big)dt + \sqrt{\frac{2}{\varepsilon}}\,dW_t \end{cases}$$

where X_t^ε are "slow" variables and Y_t^ε "fast" variables.

Expect, for $\varepsilon \to 0$, $t/\varepsilon \to +\infty$,

$$Y_t^\varepsilon \approx Y_{t/\varepsilon} \approx \int_{\mathbf{R}^n} y\,\rho_\infty(dy\,;X_t^\varepsilon), \quad Y_\cdot \text{ solving (E)},$$

so by a homogenization procedure the two-scale system should converge, as $\varepsilon \to 0$, to the averaged system

$$\dot{X}_t = -\gamma\left(X_t - \int_{\mathbf{R}^n} y\,\rho_\infty(dy\,;X_t)\right) = \nabla_x F(X_t,\gamma)$$

which is the entropic gradient descent !

The singular perturbation problem

Consider the coupled system with different time scales

$$\begin{cases} dX_t^\varepsilon = -\gamma\,(X_t^\varepsilon - Y_t^\varepsilon)\,dt \\ dY_t^\varepsilon = -\frac{1}{\varepsilon}\nabla_y\Big(f(Y_t^\varepsilon) + \frac{\gamma}{2}|X_t^\varepsilon - Y_t^\varepsilon|^2\Big)dt + \sqrt{\frac{2}{\varepsilon}}\,dW_t \end{cases}$$

where X_t^ε are "slow" variables and Y_t^ε "fast" variables.

Expect, for $\varepsilon \to 0$, $t/\varepsilon \to +\infty$,

$$Y_t^\varepsilon \approx Y_{t/\varepsilon} \approx \int_{\mathbf{R}^n} y\,\rho_\infty(dy\,;X_t^\varepsilon), \quad Y_\cdot \text{ solving (E)},$$

so by a homogenization procedure the two-scale system should converge, as $\varepsilon \to 0$, to the averaged system

$$\dot{X}_t = -\gamma\left(X_t - \int_{\mathbf{R}^n} y\,\rho_\infty(dy\,;X_t)\right) = \nabla_x F(X_t,\gamma)$$

which is the entropic gradient descent !

The algorithm: Deep Relaxation

This multiscale argument is the rationale behind an algorithm, called

Deep Relaxation, in

Chaudhari, P., Oberman, A., Osher, S., Soatto, S., Carlier, G. ,
Res. Math. Sci. 2018.

They use an Euler scheme for the two-scale system,

and implement it, with several variants, on standard computer vision datasets for training deep neural networks with the task of image classification.

Summary [Chaudhari, Osher et al. 2018]

Let $V(y,x) = \boldsymbol{f}(y) + \frac{\gamma}{2}|x-y|^2$ and the multi-scale system

$$(S\varepsilon) \qquad \begin{cases} dX_t^\varepsilon = -\nabla_x V(Y_t^\varepsilon, X_t^\varepsilon)\, dt \\ dY_t^\varepsilon = -\frac{1}{\varepsilon}\nabla_y V(Y_t^\varepsilon, X_t^\varepsilon)\, dt + \sqrt{\frac{2}{\varepsilon}}\, dW_t. \end{cases}$$

We expect that letting $\varepsilon \to 0$ yields the *deterministic averaged system*

$$(S) \qquad \frac{d\hat{X}_t}{dt} = \nabla_x F(\hat{X}_t, \gamma) = -\int_{\mathbf{R}^n} \nabla_x V(y, \hat{X}_t)\, \rho_\infty(dy\,; \hat{X}_t),$$

i.e. the gradient descent of the local entropy F corresponding to $\boldsymbol{f}$.

Our first goal: Justify the convergence $\varepsilon \to 0$.

Our second goal: Add to the problem some control variables (e.g., some tuning parameters) and prove similar convergence results.

Summary [Chaudhari, Osher et al. 2018]

Let $V(y,x) = \boldsymbol{f}(y) + \frac{\gamma}{2}|x-y|^2$ and the multi-scale system

$$(S\varepsilon) \qquad \begin{cases} dX_t^\varepsilon = -\nabla_x V(Y_t^\varepsilon, X_t^\varepsilon)\, dt \\ dY_t^\varepsilon = -\frac{1}{\varepsilon}\nabla_y V(Y_t^\varepsilon, X_t^\varepsilon)\, dt + \sqrt{\frac{2}{\varepsilon}}\, dW_t. \end{cases}$$

We expect that letting $\varepsilon \to 0$ yields the *deterministic averaged system*

$$(S) \qquad \frac{d\hat{X}_t}{dt} = \nabla_x F(\hat{X}_t, \gamma) = -\int_{\mathbf{R}^n} \nabla_x V(y, \hat{X}_t)\, \rho_\infty(dy\,; \hat{X}_t),$$

i.e. the gradient descent of the local entropy F corresponding to $\boldsymbol{f}$.

Our first goal: Justify the convergence $\varepsilon \to 0$.

Our second goal: Add to the problem some control variables (e.g., some tuning parameters) and prove similar convergence results.

Convergence of SP without controls

(A) $\quad f \in C^1(\mathbf{R}^n), \quad \nabla f$ Lipschitz, $\quad \gamma > Lip(\nabla f)$.

Theorem (1)

Assume (A), $(X^\varepsilon, Y^\varepsilon)$ solution of ($S\varepsilon$) with $X_0^\varepsilon = x$, $Y_0^\varepsilon = y$, $\hat{X}_.$ solution of (S) with $\hat{X}_0 = x$. Then, $\forall\ T > 0$, $\forall\ y \in \mathbb{R}^n$

$$\lim_{\varepsilon \to 0} \int_0^T \mathbb{E}\left[|X_s^\varepsilon - \hat{X}_s|^2\right] ds = 0,$$

$$\lim_{\varepsilon \to 0} \mathbb{E}\left[|X_T^\varepsilon - \hat{X}_T|^2\right] = 0.$$

I.o.w., the x-component of the solution of ($S\varepsilon$) converges to the solution of (S), as $\varepsilon \to 0$, for all initial positions of the y-component.

Deep relaxation with control

Motivated by Weinan E et al. 2017, we add as control parameter $u_t \in [0,1]$ the *learning rate* of the algorithm, and study the control system

$$(\mathrm{CS}\varepsilon) \qquad \begin{cases} dX_t^\varepsilon = -u_t \nabla_x V(Y_t^\varepsilon, X_t^\varepsilon)\, dt \\ dY_t^\varepsilon = -\frac{1}{\varepsilon} \nabla_y V(Y_t^\varepsilon, X_t^\varepsilon)\, dt + \sqrt{\frac{2}{\varepsilon}}\, dW_t. \end{cases}$$

with $V(y,x) = f(y) + \frac{\gamma}{2}|x-y|^2$ and the goal of **minimizing** $\mathbb{E}[f(X_T)]$.

Is there a limit control problem? and who is it?

If $\dot{X}_t^\varepsilon = u_t g_1(X_t^\varepsilon) + g_2(X_t^\varepsilon, Y_t^\varepsilon)$, i.e., u and Y separated, one expects,

as without control, $\dot{X}_t = u_t g_1(X_t) + \int_{\mathbb{R}^n} g_2(X_t, y)\rho_\infty(dy; \hat{X}_t)$,

see Kushner's book 1990, but in our case

$$\dot{X}_t^\varepsilon = -u_t \gamma(X_t^\varepsilon - Y_t^\varepsilon),$$

this does not work!

Deep relaxation with control

Motivated by Weinan E et al. 2017, we add as control parameter $u_t \in [0,1]$ the *learning rate* of the algorithm, and study the control system

$$(CS\varepsilon) \qquad \begin{cases} dX_t^\varepsilon = -u_t \nabla_x V(Y_t^\varepsilon, X_t^\varepsilon)\,dt \\ dY_t^\varepsilon = -\frac{1}{\varepsilon}\nabla_y V(Y_t^\varepsilon, X_t^\varepsilon)\,dt + \sqrt{\frac{2}{\varepsilon}}\,dW_t. \end{cases}$$

with $V(y,x) = f(y) + \frac{\gamma}{2}|x-y|^2$ and the goal of **minimizing** $\mathbb{E}[f(X_T)]$.

Is there a limit control problem? and who is it?

If $\dot{X}_t^\varepsilon = u_t g_1(X_t^\varepsilon) + g_2(X_t^\varepsilon, Y_t^\varepsilon)$, i.e., u and Y separated, one expects, as without control, $\dot{X}_t = u_t g_1(X_t) + \int_{\mathbf{R}^n} g_2(X_t, y)\rho_\infty(dy; \hat{X}_t)$,

see Kushner's book 1990, but in our case

$$\dot{X}_t^\varepsilon = -u_t\gamma(X_t^\varepsilon - Y_t^\varepsilon),$$

this does not work!

Deep relaxation with control

Motivated by Weinan E et al. 2017, we add as control parameter $u_t \in [0, 1]$ the *learning rate* of the algorithm, and study the control system

$$(\mathrm{CS}\varepsilon) \qquad \begin{cases} \mathrm{d}X_t^\varepsilon = -u_t \nabla_x V(Y_t^\varepsilon, X_t^\varepsilon)\,\mathrm{d}t \\ \mathrm{d}Y_t^\varepsilon = -\frac{1}{\varepsilon}\nabla_y V(Y_t^\varepsilon, X_t^\varepsilon)\,\mathrm{d}t + \sqrt{\frac{2}{\varepsilon}}\,\mathrm{d}W_t. \end{cases}$$

with $V(y, x) = f(y) + \frac{\gamma}{2}|x - y|^2$ and the goal of **minimizing** $\mathbb{E}[f(X_T)]$.

Is there a limit control problem? and who is it?

If $\dot{X}_t^\varepsilon = u_t g_1(X_t^\varepsilon) + g_2(X_t^\varepsilon, Y_t^\varepsilon)$, i.e., u and Y separated, one expects,

as without control, $\dot{X}_t = u_t g_1(X_t) + \int_{\mathbf{R}^n} g_2(X_t, y)\rho_\infty(\mathrm{d}y; \hat{X}_t)$,

see Kushner's book 1990, but in our case

$$\dot{X}_t^\varepsilon = -u_t\gamma(X_t^\varepsilon - Y_t^\varepsilon),$$

this does not work!

Motivated by Weinan E et al. 2017, we add as control parameter $u_t \in [0,1]$ the *learning rate* of the algorithm, and study the control system

$$(CS\varepsilon) \qquad \begin{cases} dX_t^\varepsilon = -u_t \nabla_x V(Y_t^\varepsilon, X_t^\varepsilon)\, dt \\ dY_t^\varepsilon = -\frac{1}{\varepsilon}\nabla_y V(Y_t^\varepsilon, X_t^\varepsilon)\, dt + \sqrt{\frac{2}{\varepsilon}}\, dW_t. \end{cases}$$

with $V(y,x) = f(y) + \frac{\gamma}{2}|x-y|^2$ and the goal of **minimizing** $\mathbb{E}[f(X_T)]$.

Is there a limit control problem? and who is it?

If $\dot{X}_t^\varepsilon = u_t g_1(X_t^\varepsilon) + g_2(X_t^\varepsilon, Y_t^\varepsilon)$, i.e., u and Y separated, one expects,

as without control, $\dot{X}_t = u_t g_1(X_t) + \int_{\mathbf{R}^n} g_2(X_t, y)\rho_\infty(dy; \hat{X}_t)$,

see Kushner's book 1990, but in our case

$$\dot{X}_t^\varepsilon = -u_t \gamma (X_t^\varepsilon - Y_t^\varepsilon),$$

this does not work!

Extended controls

Let $U \subseteq \mathbf{R}$ compact the set of values for the control u_t, e.g., $U = [0,1]$.
Define $U^{ex} := L^\infty(\mathbf{R}^n, U)$, and for $\nu \in U^{ex}$

$$\overline{\phi}(x,\nu) := -\int_{\mathbb{R}^n} \nu(y)\gamma(x-y)\rho^\infty(dy;x)$$

The candidate limit control system is

$$\text{(CS)} \qquad \frac{d\hat{X}_t}{dt} = \overline{\phi}(\hat{X}_t,\nu_t)\,, \quad \nu_t \in U^{ex}.$$

N.B.: if $\nu(y) \equiv u \ \forall\, y$, i.e., it is constant, then

$$\overline{\phi}(x,u) = -u\int_{\mathbb{R}^n} \gamma(x-y)\rho^\infty(dy;x) = u\nabla_x F(x,\gamma)\,,$$

i.e., the controlled Entropic Gradient Descent.

Convergence of the value functions

Define, for $\mathcal{U}$ = progressively meas.le processes in U, $(X^\varepsilon, Y^\varepsilon)$ solution of (CSε) with $X_0^\varepsilon = x$, $Y_0^\varepsilon = y$,

$$\mathcal{V}^\varepsilon(x,y) := \inf_{u.\in\mathcal{U}} \mathbb{E}[\, f(X_T^\varepsilon)\,]$$

and, for $\mathcal{U}^{ex}$ = progressively meas.le processes in U^{ex}, $\hat{X}.$ solution of (CS) with $\hat{X}_0 = x$,

$$\overline{\mathcal{V}}(x) := \inf_{\nu.\in\mathcal{U}^{ex}} f(\hat{X}_T)\,.$$

Theorem (2)

Assume (A). Then for all $T > 0$

$$\lim_{\varepsilon\to 0} \mathcal{V}^\varepsilon(x,y) = \overline{\mathcal{V}}(x) \quad \textit{locally uniformly},$$

i.e., the value functions with perturbed dynamics converge to the value of Entropic gradient descent with extended controls.

Approximation of Entropic Gradient Descent

The controlled Entropic Gradient Descent is

$$\dot{X}_t = u_t \nabla F(X_t, \gamma) \, .$$

Its value function is

$$\mathcal{V}(x) := \inf_{u_\cdot \in \mathcal{U}} f(X_T) \, , \quad X_0 = x \, .$$

Corollary

$$\lim_{\varepsilon \to 0} \mathcal{V}^\varepsilon(x, y) \leq \mathcal{V}(x),$$

i.e., the perturbed dynamics yields a value not larger than the controlled Entropic gradient descent.

This gives a further justification to the use of the Deep Relaxation algorithm for the search of the global minima of f .

Convergence of trajectories - 1

Theorem (3.a)

Assume (A) and $(X^{\varepsilon_n}, Y^{\varepsilon_n})$ *be trajectories of (CP*ε*) such that*

$$\lim_{\varepsilon_n \to 0} \int_0^T \mathbb{E}\left[|X_s^{\varepsilon_n} - \bar{x}_s|^2\right] ds = 0,$$

for a deterministic process $\bar{x}_\cdot$. *Then*

(i) $\bar{x}_\cdot$ *is a trajectory of the limit system (CS) for some control* $\nu_\cdot \in \mathcal{U}^{ex}$ *;*

(ii) if $(X^{\varepsilon_n}, Y^{\varepsilon_n})$ *are sub-optimal, i.e.,* $\mathbb{E}[f(X_T^{\varepsilon_n})] \leq \mathcal{V}^{\varepsilon_n}(x, y) + o(1)$

as $\varepsilon_n \to 0$ *, and* $\mathbb{E}\left[f(X_T^{\varepsilon_n})\right] \to f(\bar{x}_T)$ *, then*

$\bar{x}_\cdot$ *is an OPTIMAL trajectory for the limit problem, i.e.,* $f(\bar{x}_T) = \overline{\mathcal{V}}(x)$.

Convergence of trajectories - 1

Theorem (3.a)

Assume (A) and $(X^{\varepsilon_n}, Y^{\varepsilon_n})$ *be trajectories of (CP*ε*) such that*

$$\lim_{\varepsilon_n \to 0} \int_0^T \mathbb{E}\left[|X_s^{\varepsilon_n} - \bar{x}_s|^2\right] ds = 0,$$

for a deterministic process $\bar{x}_.$. *Then*

(i) $\bar{x}_.$ *is a trajectory of the limit system (CS) for some control* $\nu_. \in \mathcal{U}^{ex}$ *;*

(ii) if $(X^{\varepsilon_n}, Y^{\varepsilon_n})$ *are sub-optimal, i.e.,* $\mathbb{E}[f(X_T^{\varepsilon_n})] \le \mathcal{V}^{\varepsilon_n}(x, y) + o(1)$

as $\varepsilon_n \to 0$ *, and* $\mathbb{E}\left[f(X_T^{\varepsilon_n})\right] \to f(\bar{x}_T)$ *, then*

$\bar{x}_.$ *is an OPTIMAL trajectory for the limit problem, i.e.,* $f(\bar{x}_T) = \overline{\mathcal{V}}(x)$.

Convergence of trajectories - 1

Theorem (3.a)

Assume (A) and $(X^{\varepsilon_n}, Y^{\varepsilon_n})$ *be trajectories of (CPε) such that*

$$\lim_{\varepsilon_n \to 0} \int_0^T \mathbb{E}\left[|X_s^{\varepsilon_n} - \bar{x}_s|^2\right] ds = 0,$$

for a deterministic process $\bar{x}_.$. *Then*

(i) $\bar{x}_.$ *is a trajectory of the limit system (CS) for some control* $\nu_. \in \mathcal{U}^{ex}$ *;*

(ii) if $(X^{\varepsilon_n}, Y^{\varepsilon_n})$ *are sub-optimal, i.e.,* $\mathbb{E}[f(X_T^{\varepsilon_n})] \leq \mathcal{V}^{\varepsilon_n}(x, y) + o(1)$

as $\varepsilon_n \to 0$ *, and* $\mathbb{E}\left[f(X_T^{\varepsilon_n})\right] \to f(\bar{x}_T)$ *, then*

$\bar{x}_.$ *is an OPTIMAL trajectory for the limit problem, i.e.,* $f(\bar{x}_T) = \overline{\mathcal{V}}(x)$*.*

Convergence of trajectories - 2

Theorem (3.b)

Conversely, if $\hat{X}_\cdot$ is a trajectory of the limit system (CS), then

(i) $\exists$ a sequence $(X^{\varepsilon_n}, Y^{\varepsilon_n})$ of trajectories of (CP_ε) such that

$$\lim_{\varepsilon_n \to 0} \int_0^T \mathbb{E}\left[|X_s^{\varepsilon_n} - \hat{X}_s|^2\right] ds = 0, \quad \mathbb{E}\left[|X_T^{\varepsilon_n} - \hat{X}_T|^2\right] \to 0,$$

(ii) if, moreover, $\hat{X}_\cdot$ is optimal for the limit problem, then

$(X^{\varepsilon_n}, Y^{\varepsilon_n})$ are sub-optimal for the perturbed problem.

We also have: *if $\hat{\nu}(\cdot) \in \mathcal{U}^{ex}$ is the control corresponding to $\hat{X}_\cdot$, then the control u^{ε_n} corresponding to X^{ε_n} above also satisfies*

$$\lim_{\varepsilon_n \to 0} \int_0^T \mathbb{E}\left[\int_{\mathbb{R}^m} |u_s^{\varepsilon_n} - \hat{\nu}_s(r)|^q \, d\mu_{\hat{X}_s}(r)\right] ds = 0, \quad \forall q > 0.$$

In particular, *if $\hat{\nu}(\cdot) \in \mathcal{U}$, then no need for $\int_{\mathbb{R}^m} \ldots d\mu_{\hat{X}_s}(r)$.*

Convergence of trajectories - 2

Theorem (3.b)

Conversely, if $\hat{X}_\cdot$ is a trajectory of the limit system (CS), then

(i) $\exists$ a sequence $(X^{\varepsilon_n}, Y^{\varepsilon_n})$ of trajectories of (CP$_\varepsilon$) such that

$$\lim_{\varepsilon_n \to 0} \int_0^T \mathbb{E}\left[|X_s^{\varepsilon_n} - \hat{X}_s|^2\right] ds = 0, \quad \mathbb{E}\left[|X_T^{\varepsilon_n} - \hat{X}_T|^2\right] \to 0\,,$$

(ii) if, moreover, $\hat{X}_\cdot$ is optimal for the limit problem, then

$(X^{\varepsilon_n}, Y^{\varepsilon_n})$ are sub-optimal for the perturbed problem.

We also have: *if $\hat{\nu}(\cdot) \in \mathcal{U}^{ex}$ is the control corresponding to $\hat{X}_\cdot$, then the control u^{ε_n} corresponding to X^{ε_n} above also satisfies*

$$\lim_{\varepsilon_n \to 0} \int_0^T \mathbb{E}\left[\int_{\mathbb{R}^m} |u_s^{\varepsilon_n} - \hat{\nu}_s(r)|^q \, d\mu_{\hat{X}_s}(r)\right] ds = 0, \quad \forall\, q > 0.$$

In particular, *if $\hat{\nu}(\cdot) \in \mathcal{U}$, then no need for $\int_{\mathbb{R}^m} \dots d\mu_{\hat{X}_s}(r)$.*

Conclusions of the main results

The two-scale control system (CSε) converges to the deterministic control system

$$\text{(CS)} \qquad \frac{d\hat{X}_t}{\mathrm{d}t} = \overline{\phi}(\hat{X}_t, \nu_t)\,, \quad \nu_t \in U^{ex}$$

which is an extension of the controlled Entropic Gradient Descent, and convergence is in two senses:

- variational: ε-value function converge to limit value function,
- pathwise: ε-trajectories converge to limit trajectories, and suboptimal for (CSε) go to optimal for (CS).

Ingredients of the proofs in a general setting

We consider the more general control system of SDEs (not gradient!)

$$(S_\varepsilon) \qquad \begin{cases} dX_s^\varepsilon = \phi(X_s^\varepsilon, Y_s^\varepsilon, u_s)\, dt & X_t^\varepsilon = x \in \mathbb{R}^n \\ dY_s^\varepsilon = \dfrac{1}{\varepsilon} b(X_s^\varepsilon, Y_s^\varepsilon)\, dt + \sqrt{\dfrac{2}{\varepsilon}}\, dW_s, & Y_t^\varepsilon = y \in \mathbb{R}^m \end{cases}$$

Assumptions:

- U compact set, $\varepsilon > 0$, ϕ, b Lipschitz continuous unif. in u s.t.
- $|\phi(x, y, u)|, |b(x, y)| \le C(1 + |x| + |y|), \quad \forall\, x, y, \ \forall u \in U$
- a recurrence condition on the fast process Y:

$$\exists\, \kappa > 0 : \ (b(x, y_1) - b(x, y_2)) \cdot (y_1 - y_2) \le -\kappa\, |y_1 - y_2|^2, \ \forall\, x, y_1, y_2.$$

In the model $b(x, y) = \gamma(x - y) - \nabla f$, so the recurrence holds under condition (A).

Ingredients of the proofs in a general setting

We consider the more general control system of SDEs (not gradient!)

$$(S_\varepsilon) \qquad \begin{cases} dX_s^\varepsilon = \phi(X_s^\varepsilon, Y_s^\varepsilon, u_s)\, dt & X_t^\varepsilon = x \in \mathbb{R}^n \\ dY_s^\varepsilon = \frac{1}{\varepsilon} b(X_s^\varepsilon, Y_s^\varepsilon)\, dt + \sqrt{\frac{2}{\varepsilon}}\, dW_s, & Y_t^\varepsilon = y \in \mathbb{R}^m \end{cases}$$

Assumptions:

- U compact set, $\varepsilon > 0$, ϕ, b Lipschitz continuous unif. in u s.t.
- $|\phi(x, y, u)|, |b(x, y)| \le C(1 + |x| + |y|), \quad \forall\, x, y, \forall u \in U$
- a recurrence condition on the fast process Y:

$$\exists\, \kappa > 0 : \ (b(x, y_1) - b(x, y_2)) \cdot (y_1 - y_2) \le -\kappa\, |y_1 - y_2|^2, \ \forall\, x, y_1, y_2.$$

In the model $b(x, y) = \gamma(x - y) - \nabla f$, so the recurrence holds under condition (A).

The ε-HJB equation

Set $\mathcal{U}$ = processes with values in U progress. measurable w.r.t. W_s. Define the value function and the Hamiltonian

$$V^\varepsilon(t,x,y) := \inf_{u.\in\mathcal{U}} \mathbb{E}[\, f(X_T^\varepsilon)\,], \quad H(x,y,p) := -\min_{u\in U} \phi(x,y,u)\cdot p$$

Ass. (A) $\Rightarrow$ $|f(x)| \le K(1+|x|^2)\; \forall\, x.$

V^ε solves the Cauchy problem in $[0,T)\times\mathbb{R}^n\times\mathbb{R}^m$

$$\begin{cases} -\partial_t V^\varepsilon + H(x,y,D_x V^\varepsilon) - \dfrac{1}{\varepsilon}(b\cdot D_y V^\varepsilon + \Delta_{yy} V^\varepsilon) = 0, \\ \\ V^\varepsilon(T,x,y) = f(x), \quad \text{in } \mathbb{R}^n\times\mathbb{R}^m, \end{cases}$$

and it is the unique viscosity solution satisfying

$$|V^\varepsilon(t,x,y)| \le K(1+|x|^2+|y|^2), \quad \forall\,(t,x,y)\in[0,T]\times\mathbb{R}^n\times\mathbb{R}^m.$$

The ε-HJB equation

Set $\mathcal{U}$ = processes with values in U progress. measurable w.r.t. W_s.
Define the value function and the Hamiltonian

$$V^\varepsilon(t,x,y) := \inf_{u.\in\mathcal{U}} \mathbb{E}[f(X_T^\varepsilon)], \quad H(x,y,p) := -\min_{u\in U} \phi(x,y,u)\cdot p$$

Ass. (A) $\Rightarrow$ $|f(x)| \le K(1+|x|^2)\ \forall\, x.$

V^ε solves the Cauchy problem in $[0,T)\times\mathbb{R}^n\times\mathbb{R}^m$

$$\begin{cases} -\partial_t V^\varepsilon + H(x,y,D_x V^\varepsilon) - \dfrac{1}{\varepsilon}(b\cdot D_y V^\varepsilon + \Delta_{yy} V^\varepsilon) = 0, \\ \\ V^\varepsilon(T,x,y) = f(x), \quad \text{in } \mathbb{R}^n\times\mathbb{R}^m, \end{cases}$$

and it is the unique viscosity solution satisfying

$$|V^\varepsilon(t,x,y)| \le K(1+|x|^2+|y|^2), \quad \forall\,(t,x,y)\in[0,T]\times\mathbb{R}^n\times\mathbb{R}^m.$$

How to find the limit HJB equation

We want to show that $V^\varepsilon(t,x,y) \to v(t,x)$ as $\varepsilon \to 0$, s.t.

(Eff) $$-\partial_t v + \bar{H}(x, D_x v) = 0, \quad \text{in } [0,T) \times \mathbb{R}^n$$

Main difficulty: construct the *effective Hamiltonian* $\bar{H}$.

Try to guess it by the ansatz $V^\varepsilon(t,x,y) = v(t,x) + \varepsilon\,\chi(y) +$ l.o.t. :

$$-\partial_t v + H(x, y, D_x v) - (b \cdot D_y \chi + \Delta_{yy}\chi) = l.o.t..$$

To get the equation (Eff) for v we freeze $(\bar{x}, \bar{p})$, solve the *cell problem*

find $(c, \chi(y)) \in \mathbb{R} \times C(\mathbb{R}^m)$:

(C) $$H(\bar{x}, y, \bar{p}) - (b(\bar{x}, y) \cdot D\chi + \Delta\chi) = c$$

and finally set $\bar{H}(\bar{x}, \bar{p}) := c$. Then formally get (Eff).

(C) is an *ergodic HJB PDE*, $c =$ critical value (as in weak KAM theory), χ is called corrector.

How to find the limit HJB equation

We want to show that $V^\varepsilon(t,x,y) \to v(t,x)$ as $\varepsilon \to 0$, s.t.

(Eff) $$-\partial_t v + \bar{H}(x, D_x v) = 0, \quad \text{in } [0,T)\times\mathbb{R}^n$$

Main difficulty: construct the *effective Hamiltonian* $\bar{H}$.

Try to guess it by the ansatz $\quad V^\varepsilon(t,x,y) = v(t,x) + \varepsilon\,\chi(y) +$ l.o.t. :

$$-\partial_t v + H(x,y,D_x v) - (b\cdot D_y\chi + \Delta_{yy}\chi) = l.o.t..$$

To get the equation (Eff) for v we freeze $(\bar{x},\bar{p})$, solve the *cell problem*

find $\quad (c, \chi(y)) \in \mathbb{R}\times C(\mathbb{R}^m)$:

(C) $$H(\bar{x},y,\bar{p}) - (b(\bar{x},y)\cdot D\chi + \Delta\chi) = c$$

and finally set $\quad \bar{H}(\bar{x},\bar{p}) := c$. Then formally get (Eff).

(C) is an *ergodic HJB PDE*, $c =$ critical value (as in weak KAM theory), χ is called corrector.

How to find the limit HJB equation

We want to show that $V^\varepsilon(t,x,y) \to v(t,x)$ as $\varepsilon \to 0$, s.t.

(Eff) $$-\partial_t v + \bar{H}(x, D_x v) = 0, \quad \text{in } [0,T) \times \mathbb{R}^n$$

Main difficulty: construct the *effective Hamiltonian* $\bar{H}$.

Try to guess it by the ansatz $\quad V^\varepsilon(t,x,y) = v(t,x) + \varepsilon\,\chi(y) +$ l.o.t. :

$$-\partial_t v + H(x, y, D_x v) - (b \cdot D_y \chi + \Delta_{yy}\chi) = l.o.t..$$

To get the equation (Eff) for v we freeze $(\bar{x}, \bar{p})$, solve the *cell problem*

find $\quad (c, \chi(y)) \in \mathbb{R} \times C(\mathbb{R}^m)$:

(C) $$H(\bar{x}, y, \bar{p}) - (b(\bar{x}, y) \cdot D\chi + \Delta\chi) = c$$

and finally set $\quad \bar{H}(\bar{x}, \bar{p}) := c$. Then formally get (Eff).

(C) is an *ergodic HJB PDE*, $c =$ critical value (as in weak KAM theory), χ is called corrector.

Truncated cell problems

Classical approach: compact settings or bounded coefficients: Lions-Papanicolaou-Varadhan (1987), Kushner (1990), M.B.-Alvarez (2003-10), Borkar-Gaitsgory (2007), M.B.-Cesaroni (2010-11),....

In our problem data are unbounded in y : must change approach!

We consider a ***truncated*** $\delta-$**cell problem:**

Let D_n = ball of radius n in $\mathbb{R}^m$ and set $h(y) := H(\bar{x}, y, \bar{p})$.
Consider the Dirichlet-Poisson problem

$$\begin{cases} \delta\omega_\delta^n - (b \cdot D\omega_\delta^n + \Delta\omega_\delta^n) = -h, \text{ in } D_n \\ \omega_\delta^n = 0, \text{ on } \partial D_n. \end{cases}$$

$\Longrightarrow \quad \omega_\delta^n(y) = \mathbb{E}\left[-\int_0^{\tau_n} h(Y_y(t))e^{-\delta t}dt\right]$ is an approximate corrector,

where $\tau_n = 1^{st}$ exit time from D_n of the process

(E') $\qquad dY_y(t) = b(\bar{x}, Y_y(t))\,dt + \sqrt{2}\,dW_t, \quad Y_y(0) = y.$

Truncated cell problems

Classical approach: compact settings or bounded coefficients: Lions-Papanicolaou-Varadhan (1987), Kushner (1990), M.B.-Alvarez (2003-10), Borkar-Gaitsgory (2007), M.B.-Cesaroni (2010-11),....

In our problem data are unbounded in y : must change approach!

We consider a ***truncated*** $\delta-$**cell problem:**

Let $D_n =$ ball of radius n in $\mathbb{R}^m$ and set $h(y) := H(\bar{x}, y, \bar{p})$.
Consider the Dirichlet-Poisson problem

$$\begin{cases} \delta\omega_\delta^n - (b \cdot D\omega_\delta^n + \Delta\omega_\delta^n) = -h, \text{ in } D_n \\ \omega_\delta^n = 0, \text{ on } \partial D_n. \end{cases}$$

$\Longrightarrow \quad \omega_\delta^n(y) = \mathbb{E}\left[-\int_0^{\tau_n} h(Y_y(t))e^{-\delta t}dt\right]$ is an approximate corrector,

where $\tau_n = 1^{st}$ exit time from D_n of the process

(E') $\quad dY_y(t) = b(\bar{x}, Y_y(t))\,dt + \sqrt{2}\,dW_t, \quad Y_y(0) = y.$

Approximate correctors and $\bar{H}$

By the recurrence assumption on b the process (E') is ergodic, with unique invariant measure $\mu_{\bar{x}}$. [$\mu_{\bar{x}} = \rho_\infty(\cdot;\bar{x})$ in the model problem]

We guess the effective Hamiltonian is

$$\bar{H}(\bar{x},\bar{p}) := \int_{\mathbb{R}^m} h(y)d\mu(y) = \int_{\mathbb{R}^m} H(\bar{x},y,\bar{p})d\mu_{\bar{x}}(y).$$

Theorem

Let $\delta(n) = O\left(n^{-(4+\alpha)}\right)$ *for some* $\alpha > 0$. *Then*

$$\lim_{n\to\infty} \left| \delta(n)\,\omega^n_{\delta(n)}(y) + \bar{H} \right| = 0, \quad \textit{loc. unif. in } y.$$

Key new technical step of the proof:

fine probabilistic estimates of $\mathbb{E}\left[e^{-\delta\tau_n}\right]$, τ_n = exit time of Y from D_n.

$\Rightarrow$ The *effective HJ* Cauchy problem is

$$\begin{cases} -\partial_t v + \bar{H}(x, D_x v) = 0, & (t,x) \in (0,T)\times\mathbb{R}^n, \\ v(T,x) = f(x), & \text{in } \mathbb{R}^n. \end{cases}$$

Approximate correctors and $\bar{H}$

By the recurrence assumption on b the process (E') is ergodic, with unique invariant measure $\mu_{\bar{x}}$. [$\mu_{\bar{x}} = \rho_\infty(\cdot;\bar{x})$ in the model problem]

We guess the effective Hamiltonian is

$$\bar{H}(\bar{x},\bar{p}) := \int_{\mathbb{R}^m} h(y)d\mu(y) = \int_{\mathbb{R}^m} H(\bar{x},y,\bar{p})d\mu_{\bar{x}}(y).$$

Theorem

Let $\delta(n) = O\left(n^{-(4+\alpha)}\right)$ *for some* $\alpha > 0$. *Then*

$$\lim_{n\to\infty} \left| \delta(n)\,\omega^n_{\delta(n)}(y) \;+\; \bar{H} \right| = 0, \qquad \textit{loc. unif. in } y.$$

Key new technical step of the proof:

fine probabilistic estimates of $\mathbb{E}\left[e^{-\delta\tau_n}\right]$, $\tau_n =$ exit time of Y from D_n.

$\Rightarrow$ The *effective HJ* Cauchy problem is

$$\begin{cases} -\partial_t v + \bar{H}(x, D_x v) = 0, & (t,x) \in (0,T)\times\mathbb{R}^n, \\ v(T,x) = f(x), & \text{in } \mathbb{R}^n. \end{cases}$$

Approximate correctors and $\bar{H}$

By the recurrence assumption on b the process (E') is ergodic, with unique invariant measure $\mu_{\bar{x}}$. [$\mu_{\bar{x}} = \rho_\infty(\cdot;\bar{x})$ in the model problem]

We guess the effective Hamiltonian is

$$\bar{H}(\bar{x},\bar{p}) := \int_{\mathbb{R}^m} h(y)d\mu(y) = \int_{\mathbb{R}^m} H(\bar{x},y,\bar{p})d\mu_{\bar{x}}(y).$$

Theorem

Let $\delta(n) = O\left(n^{-(4+\alpha)}\right)$ *for some* $\alpha > 0$. *Then*

$$\lim_{n\to\infty}\left|\delta(n)\,\omega^n_{\delta(n)}(y) + \bar{H}\right| = 0, \qquad \textit{loc. unif. in } y.$$

Key new technical step of the proof:

fine probabilistic estimates of $\mathbb{E}\left[e^{-\delta\tau_n}\right]$, $\tau_n =$ exit time of Y from D_n.

$\Rightarrow$ The *effective HJ* Cauchy problem is

$$\begin{cases} -\partial_t v + \bar{H}(x, D_x v) = 0, & (t,x)\in(0,T)\times\mathbb{R}^n, \\ v(T,x) = f(x), & \text{in } \mathbb{R}^n. \end{cases}$$

A control representation of $v(t,x)$

Can prove $V^\varepsilon(t,x,y) \to v(t,x)$ = unique solution of effective problem.

Difficulty: the effective Hamiltonian $\bar{H}$ is not Bellman:

$$\bar{H}(x,p) = -\int_{\mathbf{R}^m} \min_{u\in U} \phi(x,y,u)\cdot p\, d\mu_x(y)$$

Proposition (Bellman representation of effective Ham.)

$$\bar{H}(x,p) = -\min_{\nu\in U^{ex}} \int_{\mathbf{R}^m} \phi(x,y,\nu(y))\cdot p\, d\mu_x(y)$$

which is an **exchange** formula "$\int \min = \min \int$", that uses the extended controls $U^{ex} := L^\infty(\mathbf{R}^n, U)$.

For $\nu \in U^{ex}$ take the "averaged" vector field

$$\bar{\phi}(x,\nu) := \int_{\mathbb{R}^m} \phi(x,y,\nu(y))\, d\mu_x(y)$$

Bogachev et al. 2014: $x \mapsto \mu_x$ Lip $\Rightarrow$ $\bar{\phi}$ Lip in x, unif. in ν.

A control representation of $v(t,x)$

Can prove $V^\varepsilon(t,x,y) \to v(t,x)$ = unique solution of effective problem.

Difficulty: the effective Hamiltonian $\bar{H}$ is not Bellman:

$$\bar{H}(x,p) = -\int_{\mathbf{R}^m} \min_{u\in U} \phi(x,y,u)\cdot p \, d\mu_x(y)$$

Proposition (Bellman representation of effective Ham.)

$$\bar{H}(x,p) = -\min_{\nu\in U^{ex}} \int_{\mathbf{R}^m} \phi(x,y,\nu(y))\cdot p \, d\mu_x(y)$$

which is an **exchange** formula "$\int$ min = min $\int$" , that uses the extended controls $U^{ex} := L^\infty(\mathbf{R}^n, U)$.

For $\nu \in U^{ex}$ take the "averaged" vector field

$$\bar{\phi}(x,\nu) := \int_{\mathbb{R}^m} \phi(x,y,\nu(y))\, d\mu_x(y)$$

Bogachev et al. 2014: $x \mapsto \mu_x$ Lip $\Rightarrow$ $\bar{\phi}$ Lip in x, unif. in ν.

A control representation of $v(t,x)$

Can prove $V^\varepsilon(t,x,y) \to v(t,x)$ = unique solution of effective problem.

Difficulty: the effective Hamiltonian $\bar{H}$ is not Bellman:

$$\bar{H}(x,p) = -\int_{\mathbf{R}^m} \min_{u\in U} \phi(x,y,u)\cdot p \, d\mu_x(y)$$

Proposition (Bellman representation of effective Ham.)

$$\bar{H}(x,p) = -\min_{\nu\in U^{ex}} \int_{\mathbf{R}^m} \phi(x,y,\nu(y))\cdot p \, d\mu_x(y)$$

which is an **exchange** formula " $\int \min = \min \int$ ", that uses the extended controls $U^{ex} := L^\infty(\mathbf{R}^n, U)$.

For $\nu \in U^{ex}$ take the "averaged" vector field

$$\bar{\phi}(x,\nu) := \int_{\mathbb{R}^m} \phi(x,y,\nu(y)) \, d\mu_x(y)$$

Bogachev et al. 2014: $x \mapsto \mu_x$ Lip $\Rightarrow$ $\bar{\phi}$ Lip in x unif. in ν.

Then we can consider the effective "averaged" control system

$$(\bar{S}) \qquad \begin{cases} \dfrac{d\hat{X}_t}{dt} = \bar{\phi}(\hat{X}_t, \nu_t) \\ \nu_t \in U^{ex} \text{ measurable}, \quad \text{and } \hat{X}_0 = x \in \mathbb{R}^n. \end{cases}$$

Since

$$\bar{H}(x,p) = - \min_{\nu \in U^{ex}} \bar{\phi}(x,\nu) \cdot p$$

by uniqueness of solution to the effective HJ Cauchy problem we get

$$v(t,x) = \inf f(\hat{X}_T), \quad \text{subject to } (\bar{S})$$

so v is the value function of a limit effective control problem.

Finally, we prove (as in the model problem) that

- any solution $\hat{X}$ of $(\bar{S})$ is an accumulation point in L^2 of trajectories X^ε of (S_ε),
- if a sequence X^ε solving (S_ε) converges in L^2 to a deterministic process $\bar{X}$ then

$$\frac{d\bar{X}_t}{dt} \in \overline{\text{co}}\, \bar{\phi}(\bar{x}_s, U^{ex})$$

Then we can consider the effective "averaged" control system

$$(\bar{S}) \qquad \begin{cases} \dfrac{d\hat{X}_t}{dt} = \bar{\phi}(\hat{X}_t, \nu_t) \\ \nu_t \in U^{ex} \text{ measurable}, \quad \text{and } \hat{X}_0 = x \in \mathbb{R}^n. \end{cases}$$

Since

$$\bar{H}(x,p) = -\min_{\nu \in U^{ex}} \bar{\phi}(x,\nu) \cdot p$$

by uniqueness of solution to the effective HJ Cauchy problem we get

$$v(t,x) = \inf f(\hat{X}_T), \quad \text{subject to } (\bar{S})$$

so v is the value function of a limit effective control problem.
Finally, we prove (as in the model problem) that

- any solution $\hat{X}$ of $(\bar{S})$ is an accumulation point in L^2 of trajectories X^ε of (S_ε),
- if a sequence X^ε solving (S_ε) converges in L^2 to a deterministic process $\bar{X}$ then

$$\frac{d\overline{X}_t}{dt} \in \overline{\text{co}}\, \bar{\phi}(\overline{x}_s, U^{ex})$$

Final comment and references

There are many mathematical open problems in machine learning:

"a rigorous understanding of the roots of the remarkable success of deep neural netwoks in a number of domains remains elusive".

- M. Bardi and H. Kouhkouh: Singular perturbations in stochastic optimal control with unbounded data, arXiv:2208.00655 (2022), ESAIM Control Optim. Calc. Var.
- M. Bardi and H. Kouhkouh: Deep Relaxation of Controlled Stochastic Gradient Descent via Singular Perturbations, arXiv:2209.05564 (2022).
- H. Kouhkouh: PhD thesis 2022, University of Padova.

Thanks for your attention!

Thanks to the organizers for this meeting!

We miss you Maurizio!

(San Diego 1994)

Thanks for your attention!

Thanks to the organizers for this meeting!

We miss you Maurizio!

(San Diego 1994)

Algorithm. 1 Entropy-SGD algorithm.

Input : current weights x, Langevin iterations L
Hyper-parameters: scope γ, learning rate η, SGLD step size η'

// SGLD iterations;

1 $x', \mu \leftarrow x$
2 **for** $\ell \leqslant L$ **do**
3 $\quad \Xi^\ell \leftarrow$ sample mini-batch;
4 $\quad \mathrm{d}x' \leftarrow \frac{1}{m} \sum_{i=1}^{m} \nabla_{x'} f\,(x';\ \xi_{\ell_i}) - \gamma\ (x - x')$;
5 $\quad x' \leftarrow x' - \eta'\ \mathrm{d}x' + \sqrt{\eta'}\ \epsilon\ \mathrm{N}(0, \mathrm{I})$;
6 $\quad \mu \leftarrow (1-\alpha)\mu + \alpha\ x'$;

// Update weights;

7 $x \leftarrow x - \eta\ \gamma\ (x - \mu)$

Figure: Taken from Chaudhari, Osher et al. J. Stat. Mech. (2019)

Here the process Y is denoted by x' and the process X is x.

X_t is updated in line **7**, where μ is the *average* of Y .

μ is computed by the loop in lines **2**-**6**: the *fast* process Y_t evolves (L-time) *faster* than the *slow* process X_t.