Stochastic Approximations, Diffusion Limit and Small Random Perturbations of Dynamical Systems

– a probabilistic approach to machine learning.

Wenqing Hu. [1]

1. Department of Mathematics and Statistics, Missouri S&T.

Machine Learning Background : Training DNN via minibatch Stochastic Gradient Descent (SGD).

- Deep Neural Network (DNN). Goal is to solve the following stochastic optimization problem

$$\min_{x \in \mathbb{R}^d} F(x) \equiv \frac{1}{M} \sum_{i=1}^{M} F_i(x)$$

where each component F_i corresponds to the loss function for data point $i \in \{1, ..., M\}$, and x is the vector of weights being optimized.

Machine Learning Background : Training DNN via minibatch Stochastic Gradient Descent (SGD).

- Naive Thinking : Gradient Descent (GD) updates as

$$x^{(t)} = x^{(t-1)} - \eta \left(\frac{1}{M} \sum_{i=1}^{M} \nabla F_i(x^{(t-1)}) \right) .$$

- Learning rate $\eta > 0$ is usually a small stepsize.
- Access of all gradients $\nabla F_i(x^{(t-1)})$ for $i = 1, ..., M$ is in general very expensive in large–scale machine learning problems.

Machine Learning Background : Training DNN via minibatch Stochastic Gradient Descent (SGD).

- Let B be the minibatch of prescribed size uniformly sampled from $\{1, ..., M\}$, then the objective function can be further written as the expectation of a stochastic function

$$\frac{1}{M}\sum_{i=1}^{M} F_i(x) = \mathbf{E}_B\left(\frac{1}{|B|}\sum_{i\in B} F_i(x)\right) \equiv \mathbf{E}_B F(x; B) \ .$$

- SGD updates as

$$x^{(t)} = x^{(t-1)} - \eta\left(\frac{1}{|B_t|}\sum_{i\in B_t} \nabla F_i(x^{(t-1)})\right) \ ,$$

which is the classical mini–batch version of the SGD.

Stochastic Optimization.

- Stochastic Optimization Problem

$$\min_{x \in U \subset \mathbb{R}^d} F(x) \ .$$

- $$F(x) \equiv \mathbf{E}[F(x;\zeta)] \ .$$
- Index random variable ζ follows some prescribed distribution $\mathcal{D}$.

Stochastic Optimization.

- We target at finding a local minimum point x^* of the expectation function $F(x) \equiv \mathbf{E}[F(x;\zeta)]$:

$$x^* = \arg \min_{x \in U \subset \mathbb{R}^d} \mathbf{E}[F(x;\zeta)] \ .$$

Stochastic Approximation : Stochastic Gradient Descent (SGD) Algorithm.

- Stochastic Gradient Descent (SGD) has iteration :

$$x^{(t)} = x^{(t-1)} - \eta \nabla F(x^{(t-1)}; \zeta_t) \ ,$$

where $\{\zeta_t\}$ are i.i.d. random variables that have the same distribution as $\zeta \sim \mathcal{D}$.

Closer look at SGD.

- SGD :
$$x^{(t)} = x^{(t-1)} - \eta \nabla F(x^{(t-1)}; \zeta_t) \ .$$
- $F(x) \equiv \mathbf{E} F(x; \zeta)$ in which $\zeta \sim \mathcal{D}$.
- Let
$$e_t = \nabla F(x^{(t-1)}; \zeta_t) - \nabla F(x^{(t-1)})$$
and we can rewrite the SGD as
$$x^{(t)} = x^{(t-1)} - \eta(\nabla F(x^{(t-1)}) + e_t) \ .$$

Diffusion Limit of SGD.

- Stochastic difference equation :

$$x^{(t)} - x^{(t-1)} = -\eta \nabla F(x^{(t-1)}) - \eta e_t \ .$$

- When η is small, we can approximate $x^{(t)}$ by a diffusion process X_t driven by the stochastic differential equation

$$dX_t = -\eta \nabla F(X_t) dt + \eta \sigma(X_t) dW_t \ , \ X_0 = x^{(0)} \ ,$$

where W_t is a standard Brownian motion in $\mathbb{R}^d$ and

$$\sigma(x)\sigma^T(x) = D(x) \ .$$

- The diffusion matrix (noise covariance matrix)

$$D(x) = \mathbf{E}(\nabla F(x;\zeta) - \nabla F(x))(\nabla F(x;\zeta) - \nabla F(x))^T \ .$$

Diffusion Limit of SGD : Justification.

- Slogan : Continuous Markov processes are characterized by its local statistical characteristics only in the first and second moments (conditional mean and (co)variance).
- Such an approximation has been justified in the weak sense in many classical literature. It can also be thought of as a normal deviation result (Hu–Li–Li–Liu, 2018).
- Conversely, the discrete iteration can be viewed as a numerical scheme for the diffusion limit X_t.
- In many CS literature people simply refer to the diffusion limit as the SGD algorithm.

Small random perturbations of dynamical systems : gradient flow case.

- Diffusion Limit of SGD :

$$dX_t = -\eta \nabla F(X_t)dt + \eta\sigma(X_t)dW_t \ , \ X_0 = x^{(0)} \ .$$

- Let $Y_t = X_{t/\eta}$, then

$$dY_t = -\nabla F(Y_t)dt + \sqrt{\eta}\sigma(Y_t)dW_t \ , \ Y_0 = x^{(0)} \ .$$

  (Random perturbations of the gradient flow !)

- In many CS literature people simply refer to the randomly perturbed process Y_t as the SGD algorithm.

Summary : Probabilistic Approach.

stochastic approximation

$\xrightarrow{\text{training algorithm}}$

statistical properties of learning model (convergence, generalization, ...)

$\updownarrow$ small learning rate

$\updownarrow$

diffusion limit

$\xleftarrow{\text{time change}}$

small random perturbations of dynamical systems

SGD as a randomly perturbed gradient flow.

- SGD as a randomly perturbed gradient flow :

$$dY_t = -\nabla F(Y_t)dt + \sqrt{\eta}\sigma(Y_t)dW_t \ , \ \ Y_0 = x^{(0)} \ .$$

- $$\sigma(x)\sigma^T(x) = D(x) \ .$$
- The diffusion matrix (noise covariance matrix)

$$D(x) = \mathbf{E}(\nabla F(x;\zeta) - \nabla F(x))(\nabla F(x;\zeta) - \nabla F(x))^T \ .$$

Target Problem 1 : Convergence Time.

- Assume isotropic noise $D(x) = I_d$.
- Diffusion Limit :

$$dX_t = -\eta \nabla F(X_t)dt + \eta dW_t \ , \ X_0 = x^{(0)} \ .$$

- Convergence Time = Hitting time

$$\tau^\eta = \inf\{t \geq 0 : F(X_t) \leq F(x^*) + e\}$$

  for some small $e > 0$.
- Asymptotic of $\mathbf{E}\tau^\eta$ as $\eta \to 0$?

Target Problem 1 : Convergence Time.

- Random perturbations : Let $Y_t = X_{t/\eta}$, then

$$dY_t = -\nabla F(Y_t)dt + \sqrt{\eta}dW_t \ , \ \ Y_0 = x^{(0)} \ .$$

- Hitting time

$$T^\eta = \inf\{t \geq 0 : F(Y_t) \leq F(x^*) + e\}$$

  for some small $e > 0$.

- $$\tau^\eta = \eta^{-1} T^\eta \ .$$

- Asymptotic of $\mathbf{E}\, T^\eta$ as $\eta \to 0$?

Approach to Target Problem 1 : Where is the difficulty ?

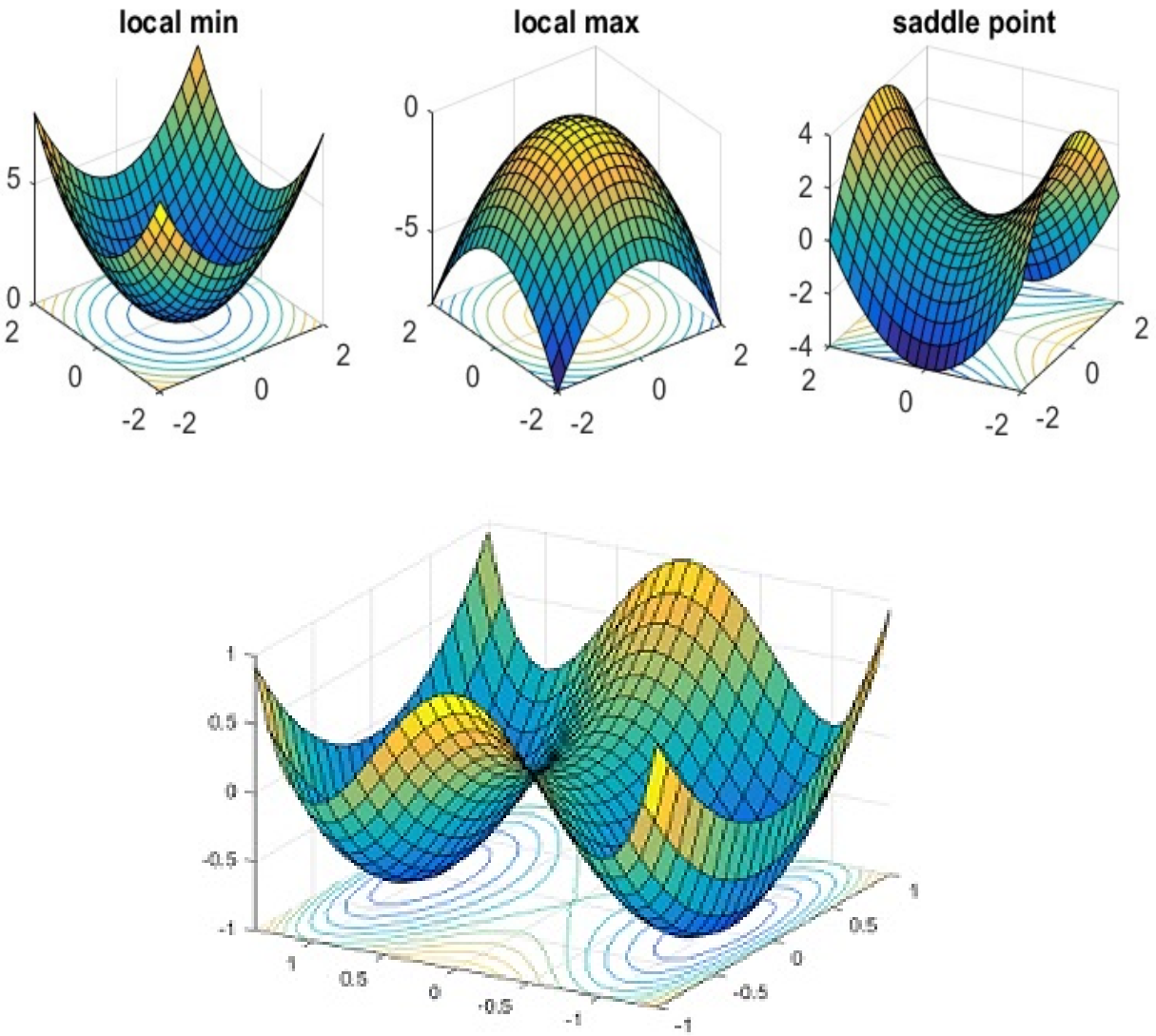

FIGURE 1 – Various critical points and the landscape of F.

Approach to Target Problem 1 : Where is the difficulty ?

- Algorithm spends most of its time near critical points of loss function F.

Approach to Target Problem 1 : Escape from a saddle point.

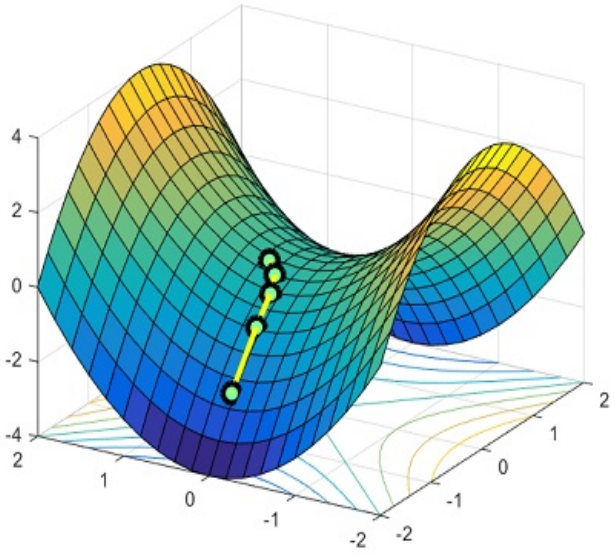

FIGURE 2 – Escape from a saddle point.

Approach to Target Problem 1 : Chain of saddle points.

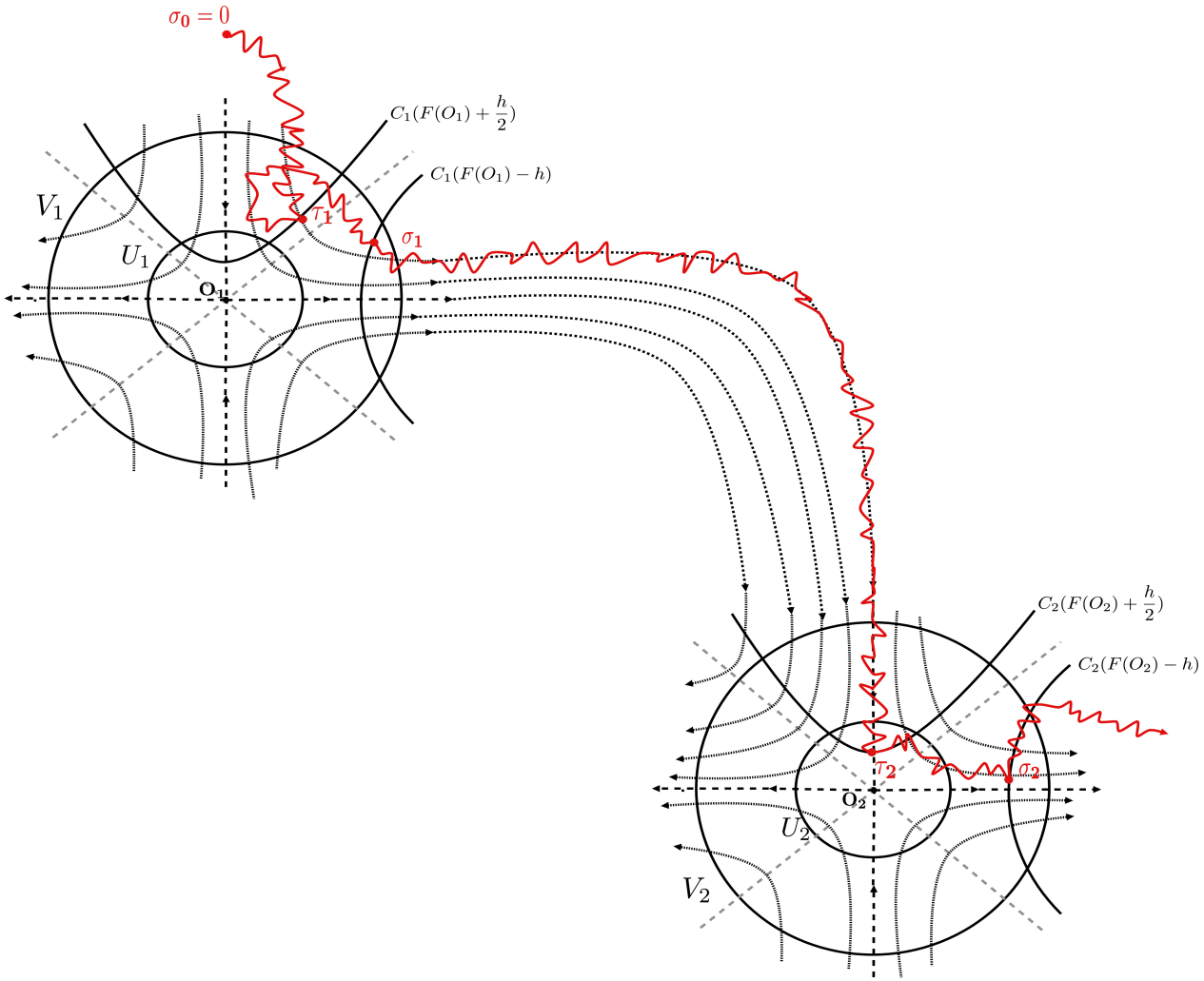

FIGURE 3 – Chain of saddle points.

Approach to Target Problem 1 : Sequence of stopping times.

- Standard Markov cycle type argument.
- But the geometry is a little different from the classical arguments for elliptic equilibriums found in Freidlin–Wentzell book.
- $\mathbf{E}\, T^{\eta} \lesssim \dfrac{k}{2\gamma_1} \ln(\eta^{-1})$ conditioned upon convergence.

Approach to Target Problem 1 : Convergence Time.

- Theorem
  (Hu-Li, 2017) (i) For any small $\rho > 0$, with probability at least $1 - \rho$, the diffusion limit X_t of SGD converges to the minimizer x^* for sufficiently small η after passing through all k saddle points O_1, ..., O_k ;
  (ii) Consider the stopping time τ^η. Then as $\eta \downarrow 0$, conditioned on the above convergence of the diffusion limit X_t of SGD, we have

$$\lim_{\eta \to 0} \frac{\mathbf{E}\tau^\eta}{\eta^{-1} \ln \eta^{-1}} \leq \frac{k}{2\gamma_1} \ .$$

Target Problem 2 : Effect of batchsize on generalization.

- It is believed that escape from "bad minima" to "good minima" is respobsible for good empirical properties under SGD training, e.g. generalization.
- DNN training : How batchsize affect the escape properties from local minimum points ?
- Large batch training v.s. small batch training.
- Large–batch methods tend to converge to sharp minimizers of the training and testing functions.
- Small–batch methods consistently converge to flat minimizers.
- $$\begin{array}{lcl} \text{Sharp minima} & \rightarrow & \text{Poorer generalization ;} \\ \text{Flat minima} & \rightarrow & \text{Better generalization .} \end{array}$$

Machine Learning Background : Training DNN via minibatch Stochastic Gradient Descent (SGD).

- Let B be the minibatch of prescribed size uniformly sampled from $\{1, ..., M\}$, then the objective function can be further written as the expectation of a stochastic function

$$\frac{1}{M}\sum_{i=1}^{M} F_i(x) = \mathbf{E}_B\left(\frac{1}{|B|}\sum_{i\in B} F_i(x)\right) \equiv \mathbf{E}_B F(x; B) \ .$$

- SGD updates as

$$x^{(t)} = x^{(t-1)} - \eta\left(\frac{1}{|B_t|}\sum_{i\in B_t} \nabla F_i(x^{(t-1)})\right) \ ,$$

which is the classical mini–batch version of the SGD.

Approach to Target Problem 2 : Relationship between batchsize and the diffusion matrix $D(x)$.

- Relationship between batchsize and the diffusion matrix $D(x)$ (Hu-Li-Li-Liu, 2018).
- $$D(x) = \left(\frac{1}{|B|} - \frac{1}{M} \right) D_0(x) \ ,$$

  where

  $$D_0(x) = \frac{1}{M-1} \sum_{i=1}^{M} (\nabla F_i(x) - \nabla F(x))(\nabla F_i(x) - \nabla F(x))^T \ .$$
- Naively, the larger $D(x)$ is, the easier to escape due to injection of the noise.

SGD as a randomly perturbed gradient flow.

- SGD as a randomly perturbed gradient flow :

$$dY_t = -\nabla F(Y_t)dt + \sqrt{\eta}\sigma(Y_t)dW_t \ , \ \ Y_0 = x^{(0)} \ .$$

- $$\sigma(x)\sigma^T(x) = D(x) \ .$$
- The diffusion matrix (noise covariance matrix)

$$D(x) = \mathbf{E}(\nabla F(x;\zeta) - \nabla F(x))(\nabla F(x;\zeta) - \nabla F(x))^T \ .$$

Approach to Target Problem 2 : Escape from basin of attractors via Large Deviations Theory.

- Theorem
  (Hu-Li-Li-Liu, 2018) Let U be a basin of attractor such that $O \subset U$ is the only minimum of $F(x)$ and assume that the boundary ∂U of the domain U is smooth so that $\nabla F(x)$, $x \in \partial U$ points to the interior of the boundary of U, then for $x_0 \in U$ we have the following two asymptotic

$$P(x_0, \partial U) \asymp \exp\left(-\frac{1}{\eta} \min_{x \in \partial U} \phi_{loc}^{QP}(x; O)\right) ,$$

and

$$\mathbf{E}\tau(x_0, \partial U) \asymp \frac{1}{\eta} \exp\left(\frac{1}{\eta} \min_{x \in \partial U} \phi_{loc}^{QP}(x; O)\right) .$$

Approach to Target Problem 2 : The quasi–potential.

- The local quasi–potential

$$\phi^{\mathrm{QP}}_{\mathrm{loc}}(x;x_0) = \inf_{T>0} \inf_{\psi(0)=x_0,\psi(T)=x} S_{0T}(\psi) \ .$$

- Action functional (Large Deviations Rate Function) :

$$S_{0T}(\psi) = \begin{cases} \dfrac{1}{2}\displaystyle\int_0^T (\dot{\psi}_t + \nabla F(\psi_t))^T D^{-1}(\psi_t)(\dot{\psi}_t + \nabla F(\psi_t))dt \ , \\ \qquad\qquad\qquad\qquad \text{if } \psi_t \text{ is abs. cont. } t \in [0,T] \ ; \\ +\infty \ , \\ \qquad\qquad\qquad\qquad \text{otherwise} \ . \end{cases}$$

Target Problem 3 : Algorithm's implicit selection of specific local minima.

- Implicit Regularization : It is widely believed that SGD is an implicit regularizer which helps itself to search for a local minimum that is easy to generalize.
- How to understand this phenomenon ?

Approach to Target Problem 3 : Structure of diffusion matrix $D(x)$.

- Covariance structure (diffusion matrix) $D(x)$ depends implicitly on the model architecture :

$$D(x) = \mathbf{E}(\nabla F(x;\zeta) - \nabla F(x))(\nabla F(x;\zeta) - \nabla F(x))^T \ .$$

- Anisotropic Noise v.s. Isotropic Noise.
- $D(x)$ may have very large condition number !

SGD as a randomly perturbed gradient flow.

- SGD as a randomly perturbed gradient flow :

$$dY_t = -\nabla F(Y_t)dt + \sqrt{\eta}\sigma(Y_t)dW_t \ , \ \ Y_0 = x^{(0)} \ .$$

- $$\sigma(x)\sigma^T(x) = D(x) \ .$$
- The diffusion matrix (noise covariance matrix)

$$D(x) = \mathbf{E}(\nabla F(x;\zeta) - \nabla F(x))(\nabla F(x;\zeta) - \nabla F(x))^T \ .$$

Approach to Target Problem 2 : Escape from basin of attractors via Large Deviations Theory.

- Theorem
  (Hu-Li-Li-Liu, 2018) Let U be a basin of attractor such that $O \subset U$ is the only minimum of $F(x)$ and assume that the boundary ∂U of the domain U is smooth so that $\nabla F(x)$, $x \in \partial U$ points to the interior of the boundary of U, then for $x_0 \in U$ we have the following two asymptotic

$$P(x_0, \partial U) \asymp \exp\left(-\frac{1}{\eta} \min_{x \in \partial U} \phi_{loc}^{QP}(x; O)\right) ,$$

and

$$\mathbf{E}\tau(x_0, \partial U) \asymp \frac{1}{\eta} \exp\left(\frac{1}{\eta} \min_{x \in \partial U} \phi_{loc}^{QP}(x; O)\right) .$$

Approach to Target Problem 2 : The quasi–potential.

- The local quasi–potential

$$\phi^{\mathrm{QP}}_{\mathrm{loc}}(x;x_0) = \inf_{T>0} \inf_{\psi(0)=x_0,\psi(T)=x} S_{0T}(\psi) \ .$$

- Action functional (Large Deviations Rate Function) :

$$S_{0T}(\psi) = \begin{cases} \dfrac{1}{2}\displaystyle\int_0^T (\dot{\psi}_t + \nabla F(\psi_t))^T D^{-1}(\psi_t)(\dot{\psi}_t + \nabla F(\psi_t))dt \ , \\ \qquad\qquad\qquad\qquad \text{if } \psi_t \text{ is abs. cont. } t \in [0,T] \ ; \\ +\infty \ , \\ \qquad\qquad\qquad\qquad \text{otherwise} \ . \end{cases}$$

Approach to Target Problem 3 : Relation between the quasi–potential and the diffusion matrix $D(x)$.

- Theorem
  (Hu-Zhu-Xiong-Huan, 2019) The local quasi–potential $\phi_{loc}^{QP}(x;x_0)$ is a solution to the Hamilton–Jacobi equation

$$\frac{1}{2}\left(\nabla\phi_{loc}^{QP}(x;x_0)\right)^T D(x)\nabla\phi_{loc}^{QP}(x;x_0) - \nabla F(x)\cdot\nabla\phi_{loc}^{QP}(x;x_0) = 0 \ ,$$

with boundary condition

$$\phi_{loc}^{QP}(O;x_0) = 0 \ .$$

Approach to Target Problem 3 : Escape from local minimum point : Isotropic Noise v.s. Anisotropic Noise.

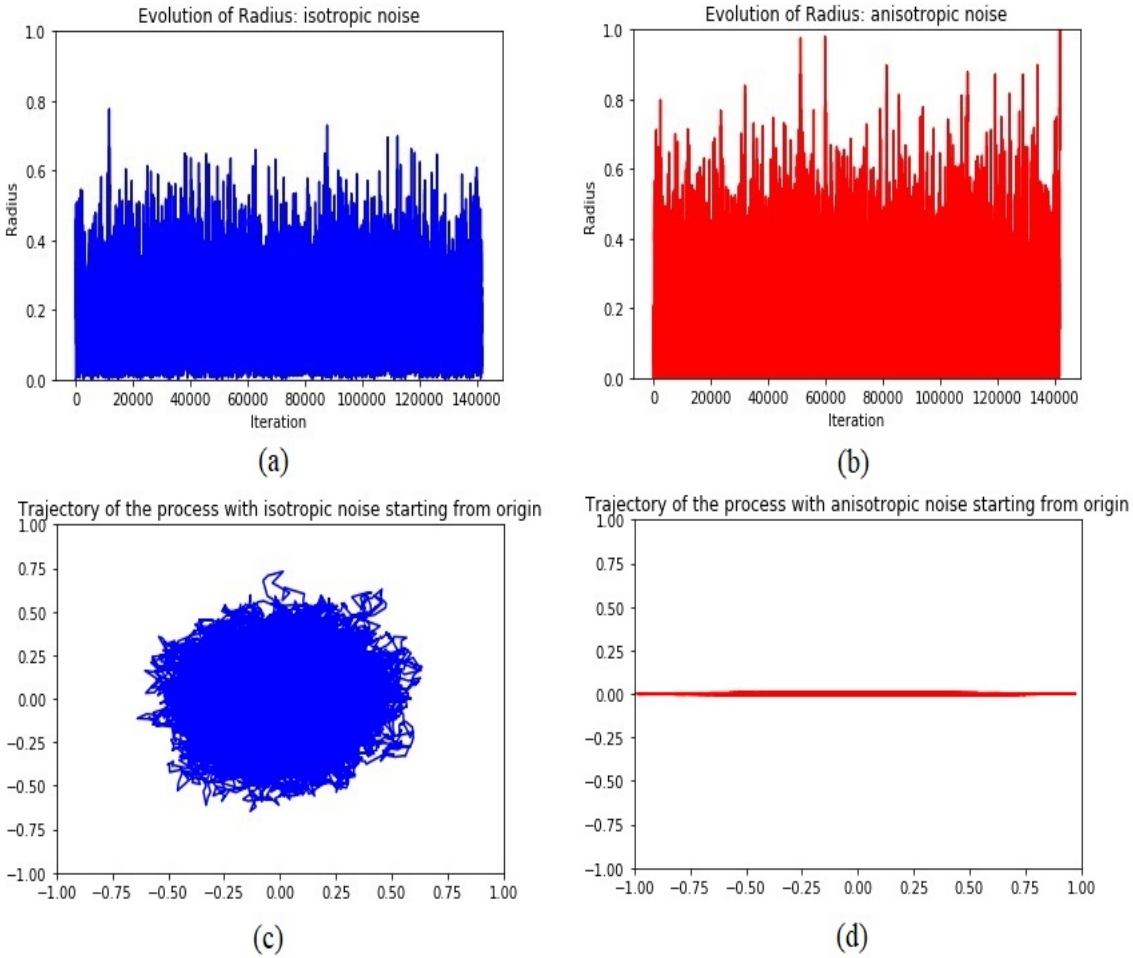

FIGURE 4 – Escape from local minimum point : Isotropic Noise v.s. Anisotropic Noise.

Target Problem 3 : Algorithm's implicit selection of specific local minima.

- Suppose that $F(x)$ is non–convex and admits several local minimum points.
- Does SGD always select the global minimum of $F(x)$?

SGD as a randomly perturbed gradient flow.

- SGD as a randomly perturbed gradient flow :

$$dY_t = -\nabla F(Y_t)dt + \sqrt{\eta}\sigma(Y_t)dW_t \ , \ \ Y_0 = x^{(0)} \ .$$

- $$\sigma(x)\sigma^T(x) = D(x) \ .$$
- The diffusion matrix (noise covariance matrix)

$$D(x) = \mathbf{E}(\nabla F(x;\zeta) - \nabla F(x))(\nabla F(x;\zeta) - \nabla F(x))^T \ .$$

Approach to Target Problem 3 : Variational Inference.

- Variational Inference : Invariant density

$$\rho^{\mathrm{SS}}(x) = \frac{1}{Z(\eta)} \exp\left(-\frac{2}{\eta}\Phi(x)\right) \ .$$

- SGD selects the global minimum of $\Phi(x)$: $\rho(x,t) \to \rho^{\mathrm{SS}}(x)$ as $t \to \infty$.
- If diffusion matrix $D(x) = I_d$ is isotropic, then $\Phi(x) \propto F(x)$.
- Anisotropic noise is common !
- So in general $\Phi(x) \neq F(x)$.

Approach to Target Problem 3 : Variational Inference via Large Deviation Theory.

- Large Deviation Theory provides another way to express the invariant density

$$\rho^{\mathsf{SS}}(x) \asymp \exp\left(-\frac{1}{\eta}\phi^{\mathsf{QP}}(x)\right) \ .$$

- (global) quasi–potential : $\phi^{\mathsf{QP}}(x)$.

Approach to Target Problem 3 : Variational Inference via Large Deviation Theory.

- Theorem
  (Hu-Zhu-Xiong-Huan, 2019) SGD selects the global minimum x^* of $\phi^{QP}(x)$ such that $\phi^{QP}(x^*) = 0$.

Approach to Target Problem 3 : Global quasi-potential.

- The (global) quasi-potential $\phi^{\mathrm{QP}}(x)$ can be calculated from the local quasi–potential $\phi^{\mathrm{QP}}_{\mathrm{loc}}(x; x_0)$.
- We have seen that the local quasi–potential $\phi^{\mathrm{QP}}_{\mathrm{loc}}(x; x_0)$ depends on the diffusion matrix $D(x)$ via the Hamilton-Jacobi equation.
- Markov chain on local minimum points (Freidlin–Wentzell theory).

Approach to Target Problem 3 : Implicit regularization via covariance structure.

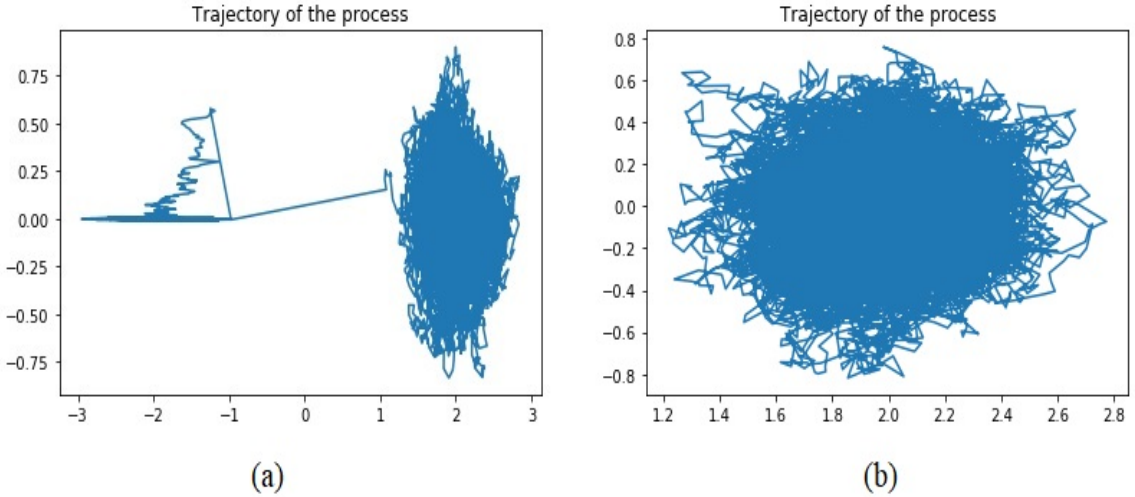

FIGURE 5 – Implicit regularization via covariance structure. The loss function is symmetric with respect to two local minima $(-2, 0)$ and $(2, 0)$. Left : Process starts from $(-2, 0)$, anisotropic noise in the potential well ; Right : Process starts from $(2, 0)$, isotropic noise in the potential well. SGD tends to select $(2, 0)$.

Summary : Probabilistic Approach.

stochastic approximation

$\xrightarrow{\text{training algorithm}}$

statistical properties of learning model (convergence, generalization, ...)

$\updownarrow$ small learning rate

$\updownarrow$

diffusion limit

$\xleftarrow{\text{time change}}$

small random perturbations of dynamical systems

Stochastic Approximation : Stochastic heavy–ball method.

- $$\min_{x \in \mathbb{R}^d} f(x) \ .$$
- "heavy ball method" (B.Polyak 1964)

$$\begin{cases} x_t & = x_{t-1} + \sqrt{s} v_t \ , \\ v_t & = (1 - \alpha\sqrt{s}) v_{t-1} - \sqrt{s} \nabla_X f\left(x_{t-1}\right) \ , \\ x_0, v_0 & \in \mathbb{R}^d \ . \end{cases}$$

- Stochastic heavy ball method :

$$\begin{cases} x_t & = x_{t-1} + \sqrt{s} v_t \ , \\ v_t & = (1 - \alpha\sqrt{s}) v_{t-1} - \sqrt{s} \nabla_X f\left(x_{t-1}; \zeta_t\right) \ , \\ x_0, v_0 & \in \mathbb{R}^d \ . \end{cases}$$

Diffusion Limit of Stochastic heavy–ball method.

- Diffusion limit of the stochastic heavy ball method :

$$\begin{cases} dx(t) & = \sqrt{s}v(t)dt \ , \\ dv(t) & = -\sqrt{s}\alpha v(t)dt - \sqrt{s}\nabla_X f(x(t)) + \sqrt{s}\sigma(x(t))dW_t \ , \\ x(0), v(0) & \in \mathbb{R}^d \ . \end{cases}$$

Small random perturbations of dynamical systems : dissipative Hamiltonian system.

- Small random perturbations of dissipated Hamiltonian system :

$$\begin{cases} dX_t^\varepsilon & = V_t^\varepsilon dt; \\ dV_t^\varepsilon & = (-\alpha V_t^\varepsilon - \nabla_X f(X_t^\varepsilon))dt + \varepsilon\sigma(X_t^\varepsilon)dW_t, \\ X_0^\varepsilon, V_0^\varepsilon & \in \mathbb{R}^d \ . \end{cases}$$

- Hu-Li-Su, 2017 and work in progress.

Many further interesting directions at the interplay of probability and data science ...

- Escape from saddle points for anisotropic noise.
- Can the program on Hamilton–Jacobi equation be carried to specific deep neural network structures ?
- Empirical Risk v.s. Population Risk : How does this affect SGD's variational inference ? Relation with Generalization.
- Many more applied problems in collaborations with CS/ECE/ORIE people...

Thank you for your attention !