Imperial College London

TO INTERACT OR NOT?

The benefits of interacting particles: convergence properties and variance reduction

Anastasia Borovykh, Panos Parpas, Nikolas Kantas and Greg Pavliotis

Machine Learning in Quantitative Finance and Risk Management, CWI July 2020

OPTIMIZATION: THE CHALLENGES IN DEEP LEARNING

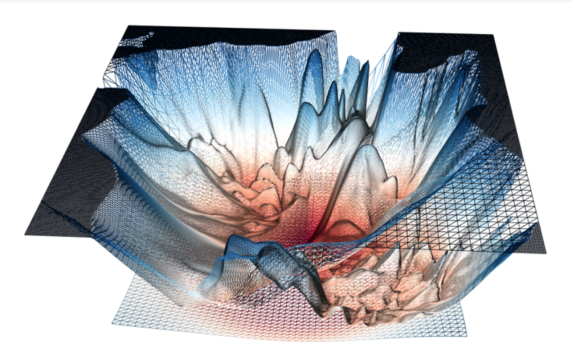

Figure: A highly nonconvex loss surface; as is common in deep neural nets [1]

[1] From `https://www.telesens.co/2019/01/16/neural-network-loss-visualization`

THE GOAL OF A GOOD OPTIMIZATION ALGORITHM

- Controllable through hyperparameters
- Escape local minima and saddle points
- Converge close to the optimum

OVERVIEW

- The optimization objective
- Classical optimization schemes and their convergence results
- The challenge with stochastic optimization
- Interacting particles: the setup
- Interacting particles: convergence properties
- Numerical examples
- Further ideas

CONSTRAINED OPTIMIZATION: THE OBJECTIVE

- Consider a generic optimization problem of the form:

$$\min_{x \in \mathcal{X}} f(x).$$

- $\mathcal{X} \in \mathbb{R}^d$ is a closed convex set describing the constraints.
- f is the objective function, taken to be L-Lipschitz and (strongly) convex.
- We are looking for the minimizer x^*

PROJECTED GRADIENT DESCENT

- One way of finding the solution is using,

$$x_{k+1} = \Pi_{\mathcal{X}}(x_k - \eta \nabla f(x_k)),$$

  where k denotes discrete time, η is the learning rate, and the projection is defined through the Euclidean norm,

$$\Pi_{\mathcal{X}}(y) = \arg\min_{x \in \mathcal{X}} ||y - x||_2^2,$$

- Drawback: tied to the Euclidean norm can lead to slow convergence for large dimension d.
- If $\sup_{x,x' \in \mathcal{X}} ||x - x'||_\infty \leq 1$ implies $\sup_{x,x' \in \mathcal{X}} ||x - x'||_2 \ < 2\sqrt{d}$ it converges at rate $\sqrt{d/t}$.

MIRROR DESCENT

- A generalization of projected gradient descent (GD),

$$x_{k+1} = \arg\min_{x \in \mathcal{X}} D_{\boldsymbol{\Phi}}(x, y_{k+1} = \nabla \Phi^*(z_{k+1}),$$

$$z_{k+1} = \nabla \Phi(x_k) - \eta \nabla f(x_k).$$

- Here $\Phi : \mathcal{X} \rightarrow \mathbb{R}^d$ is the mirror map, mapping from the constrained set to an unconstrained one.
- Its convex conjugate is $\nabla \Phi^*(z) := \arg\min_{x \in \mathcal{X}}(\Phi(x) - z^T x)$.
- The Bregman divergence is $D_{\boldsymbol{\Phi}}(x, y) = \Phi(x) - \Phi(y) - \nabla \Phi(y)^T (x - y)$.

MIRROR DESCENT

- The point x_k is mapped into its dual space mirror image $z_k = \nabla\Phi(x_k)$.
- This is updated by the negative gradient step to z_{k+1}.
- Then it is mapped back $x_{k+1} = \nabla\Phi^*(z_{k+1})$ into primal space $\mathcal{X}$.
- Note: when $\Phi(x) = \frac{1}{2}||x||_2^2$ we get back to projected GD.
- Benefits: The mapping is now done using Bregman divergence. By choosing Φ in such a way that $\sup_{x,x'} \sqrt{2D_\Phi(x,x')}$ is dimension-independent, fast convergence is obtained [2]

[2] A. BECK, First-order methods in optimization

CONTINUOUS MIRROR DESCENT

- The continuous version of mirror descent is given by,

$$\begin{aligned} dz_t &= -\nabla f(x_t)dt, \\ x_t &= \nabla \Phi^*(z_t). \end{aligned}$$

- Using Euler discretization with $\Delta t = \eta$ we obtain the discrete version.

CONTINUOUS MIRROR DESCENT: CONVERGENCE PROPERTIES

- Converge of MD for a convex objective
- Assume f is convex. Then,

$$\frac{1}{T}\int_0^T (f(x_t) - f(x^*))dt \leq \frac{D^2_{\Phi,\mathcal{X}}}{2T},$$

  where $D_{\Phi,\mathcal{X}} = \sup_{x,x'} \sqrt{2D_\Phi(x,x')}$.
- Therefore, if we choose Φ such that it ‘adapts well’ to constraint set $\mathcal{X}$ convergence can be faster than projected GD.
- In the case of a strongly convex f the convergence speed can be increased to an exponential rate.

STOCHASTIC MIRROR DESCENT

- The gradient estimate can contain noise, e.g. when computed over batches as is common in machine learning
- Under certain assumptions on the noise, we obtain stochastic mirror descent. In continuous time it is given by,

$$
\begin{aligned}
dz_t &= -\nabla f(x_t)dt + \sigma dB_t, \\
x_t &= \nabla \Phi^*(z_t),
\end{aligned}
$$

where B_t is a Brownian motion and σ determines the noise variance and $\nabla \Phi^* : \mathcal{R}^d \to \mathcal{X}$ is a projection operator.

CONVERGENCE OF STOCHASTIC MIRROR DESCENT

- **Converge of SMD for a convex objective**
- Assume f is convex. Then,

$$\mathbb{E}\left[\frac{1}{T}\int_0^T (f(x_t) - f(x^*))dt\right] \leq \frac{1}{2T}D^2_{\Phi,\mathcal{X}} + \frac{1}{2}\sigma^2||\Delta\Phi^*||_\infty.$$

- Challenge: the gap to optimality is bounded from above by a quantity proportional to noise variance σ^2...

COMMON VARIANCE REDUCTION TECHNIQUES

- When the gradient estimate contains noise, convergence is to a neighborhood of the optimum. The size of the neighborhood is controlled by σ.
- How to converge closer to the optimum?
- Increase batch size or decrease noise variance σ
- Decrease learning rate [3]
- Importance sampling [4]
- Variance reduction techniques [5], [6]

[3] P. MERTIKOPOULOS AND M. STAUDIGL, On the convergence of gradient-like flows with noisy gradient input

[4] D. NEEDELL, R. WARD, AND N. SREBRO, Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm

[5] R. JOHNSON AND T. ZHANG, Accelerating stochastic gradient descent using predictive variance reduction

[6] A. DEFAZIO, F. BACH, AND S. LACOSTE-JULIEN, SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives

CONSTRAINED OPTIMIZATION WITH INTERACTING PARTICLES

- We consider an alternative to such variance reduction techniques.
- Instead of using independent particles, one could consider interacting particles (ISMD) [7], [8], [9]

$$dz_t^i = -\nabla f(x_t^i)dt + \theta \sum_{j=1}^{N} A_{ij}(z_t^j - z_t^i) + \sigma dB_t^i,$$

  for $i = 1, ..., N$.
- Here A is the interaction matrix taken to be doubly stochastic. $A_{ij} = 1$ if two particles interact, e.g. exchange values.
- The parameter θ controls the interaction strength.

[7] M. RAGINSKY AND J. BOUVRIE, Continuous-time stochastic mirror descent on a network: Variance reduction, consensus, convergence

[8] J.C.DUCHI, A.AGARWAL, AND M.J.WAINWRIGHT, Dual averaging for distributed optimization: Convergence analysis and network scaling

[9] P. LIN, W. REN, AND J. A. FARRELL, Distributed continuous-time optimization: nonuniform gradient gains, finite-time convergence, and convex constraint set

ISMD: A USEFUL REPARAMETRIZATION

- Let $\mathbf{z}_t = ((z_t^1)^T, ..., (z_t^N)^T)^T$.
- Define the graph Laplacian as $L := \mathrm{Diag}(A\mathbf{1}_N) - A$, and let $\mathcal{L} := L \otimes I_d$, where $\otimes$ is the Kronecker product.
- Then we have,

$$d\mathbf{z}_t = (-\nabla \mathbf{V}(\mathbf{z}_t) - \mathcal{L}\mathbf{z}_t)\, dt + \sigma d\mathbf{B}_t,$$

  where $\mathbf{B}_t := ((B_t^1)^T, ..., (B_t^N)^T)^T$ is the stacked variable of Brownian motions and $\nabla \mathbf{V}(\mathbf{z}_t) = (\nabla \mathcal{V}(z_t^1)^T, ..., \nabla \mathcal{V}(z_t^N)^T)^T$.
- We have taken $\nabla \mathcal{V}(z) = \nabla f \circ \nabla \Phi^*(z)$.
- By the properties of the Laplacian, it has eigenvalues $\lambda_0 = 0 < \underline{\lambda} \leq \lambda_2 \leq ... \leq \lambda_N$.

INTERACTING PARTICLES: WHY IS IT USEFUL?

- Convergence of ISMD for a convex objective
- Let f be a convex function. Define $\tilde{z}_t^i = z_t^i - \frac{1}{N}\sum_{i=1}^N z_t^i$. Then we have,

$$\frac{1}{T}\int_0^T \mathbb{E}[(f(x_t^i) - f(x^*))]dt \le \frac{1}{2T}D^2_{\Phi,\mathcal{X}} + \frac{\sigma^2}{2N}||\Delta\Phi^*||_\infty$$
$$+ \int_0^T \frac{L}{\mu T}\mathbb{E}\left[||\tilde{z}_t^i||_*\right]dt + \int_0^T \frac{2L}{\mu NT}\sum_{i=1}^N \mathbb{E}\left[||\tilde{z}_t^i||_*\right]dt.$$

INTERACTING PARTICLES: WHY IS IT USEFUL?

- $\frac{1}{2T}D^2_{\Phi,\mathcal{X}}$: the standard optimization error giving linear in time convergence,
- $\frac{\sigma^2}{2N}||\Delta\Phi^*||_\infty$: variance is decreased by a factor of N,
- $\int_0^T \frac{L}{\mu T}\mathbb{E}\left[||\tilde{z}_t^i||_*\right]$ and $\int_0^T \frac{2L}{\mu NT}\sum_{i=1}^N \mathbb{E}\left[||\tilde{z}_t^i||_*\right]$ dt measure deviation from the particle average.
- If, fluctuation term $\tilde{z}_t^i$ is bounded and non-increasing with N, variance is reduced!

BOUNDING THE FLUCTUATION TERM

- Luckily, fluctuation is bounded under certain assumptions.
- We have, for a κ-strongly convex function f,

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}||\tilde{z}_t^i||_*^2\right] \leq e^{-\theta(\kappa+\underline{\lambda})t}C + \frac{dK}{\theta(\kappa+\underline{\lambda})}\sigma^2\frac{N-1}{N}.$$

- For a sufficiently large $\theta(\kappa+\underline{\lambda})$ the interaction is controlled.
- Strong convexity κ plays a role.
- Interaction strength θ plays a role.
- Connectivity of the particles plays a role through $\underline{\lambda}$.

THE BENEFITS OF ISMD

- For a large amount of particles N, if
    - function if sufficiently strongly convex,
    - or interaction strength is high enough,
    - and interaction between particles is dense enough,
- variance is reduced compared to running just one particle!

ANOTHER PERSPECTIVE: LOOKING AT IT FROM A SAMPLING VIEW

- If we sample long enough using ISMD our samples will be from the invariant distribution,

$$\begin{aligned}\eta_{\infty}^{N}(d\mathbf{z}) &= \frac{1}{Z_N}\exp\left(-\frac{2}{\sigma^2}\left(\sum_{i=1}^{N}\mathcal{V}(z^i)+\frac{\theta}{2}\mathbf{z}^T\mathcal{L}\mathbf{z}\right)\right)d\mathbf{z}\\ &= \frac{1}{Z_N}\exp\left(-\mathcal{W}(\mathbf{z})\right)d\mathbf{z},\end{aligned}$$

  where Z_N is the normalization constant.

ANOTHER PERSPECTIVE: LOOKING AT IT FROM A SAMPLING VIEW

- Finding the mode of η_∞^N is equivalent to solving the following optimization problem:

$$z^* = \arg\min_z \mathcal{W}(z).$$

- Observe that this is the same as,

$$z^* = \arg\min_z \mathcal{V}(z),$$

  where under the right assumptions z^* in dual space gives us the optimum x^* in primal space.
- Therefore, if the samples have converged to samples from the invariant measure, the samples lie around the minimizer!

ANOTHER PERSPECTIVE: LOOKING AT IT FROM A SAMPLING VIEW

- Studying the convergence rates from a sampling perspective, i.e. the speed of convergence of the samples z_t^i to samples from the invariant distribution can give insight into how close we are to the minimum.
- Using Bakry-Emery theory, if we can show that

$$\mathrm{Hess}(\mathcal{W}) \succeq \rho I_d,$$

then this implies,

$$||\eta_t^N - \eta_\infty^N||_{TV} \leq Ke^{-2\rho t},$$

CONVERGENCE TO THE INVARIANT MEASURE

- We can show the curvature assumption holds with with $\rho = \frac{\sigma^2}{2}(\kappa + \frac{\lambda}{2})$ where κ is the curvature of our objective function and λ is the second-lowest eigenvalue of the Laplacian matrix of the interactions.
- Therefore we have an exponential convergence to the invariant measure,

$$||\eta_t^N - \eta_\infty^N||_{TV} \leq Ke^{-2\rho t}.$$

DISTANCE TO OPTIMALITY

- Studying the form of the invariant measure can provide us with insight into how far we are from the optimum and how close the particles are.
- Remember, the mode of the invariant measure is our optimal point.
- Measuring the distance between the expected value of the samples and the mode can give us insight into the distance to optimality,

$$\mathbb{E}_{\eta_\infty^{\mathsf{N}}}(\mathcal{W}) - \mathcal{W}(\mathsf{z}^*) \leq \frac{\sigma^2}{2}\left(\frac{2\mathsf{dN}}{\rho} - \frac{1}{2}\log\left(\frac{\sigma^2}{2\mathsf{L}_\mathsf{N}}\right)\right).$$

NUMERICAL EXAMPLES: AN ILL-CONDITIONED PROBLEM

- Consider the problem,

$$\min_{x \in \mathcal{X}} ||Wx - b||_2^2, \tag{1}$$

  where $\mathcal{X} = \Delta_n$, the unit simplex, $W \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$.
- If we let W have a high condition number, the problem is ill-conditioned.

NUMERICAL EXAMPLES: AN ILL-CONDITIONED PROBLEM

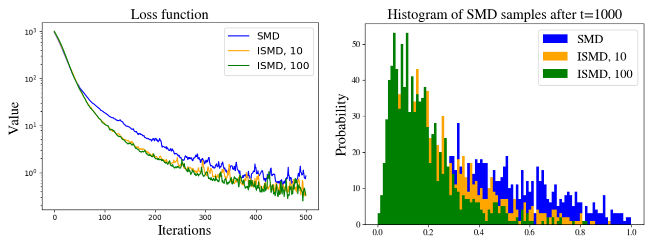

Figure: A comparison between the initial convergence of SMD and ISMD for condition number 100 (L) and a histogram (R). We observe a speedup in convergence using interacting particles and a smaller variance.

TRAFFIC ASSIGNMENT PROBLEM

- The objective of the traffic assignment problem is to compute the optimal path between two nodes in a graph.
- This problem is a convex optimization problem with a simplex constraint and therefore fits our framework.

TRAFFIC ASSIGNMENT PROBLEM

- The setup is: take a directed multi-graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. An origin-destination pair (o, d) sends λ units of traffic from o to d via a set of paths $p \in \mathcal{P}$ using edges in $\mathcal{G}$.
- The set of feasible routes is given by $\mathcal{X} = \Big((x_p)_{p\in\mathcal{P}} : x_p \geq 0 \text{ and } \sum_{p\in\mathcal{P}} x_p = \lambda\Big)$.
- The delay along a path is $c_p(x) = \sum_{e\in p} c_e(w_e)$ with w_e the load on an edge and the delay is $c_e(w_e)$.
- The average delay is then $C(x) = \sum_{p\in\mathcal{P}} x_p c_p(x)$.
- The objective is to find an optimum routing flow $x^* \in \arg\min_{x\in\mathcal{X}} C(x)$.

TRAFFIC ASSIGNMENT PROBLEM

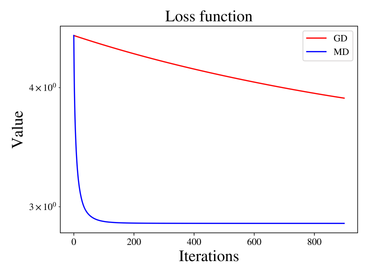

Figure: A comparison between MD and GD. GD with Euclidean projections is considerably slower than MD for this problem because the solution is sparse. IMD performs the same as MD for deterministic problems.

TRAFFIC ASSIGNMENT PROBLEM

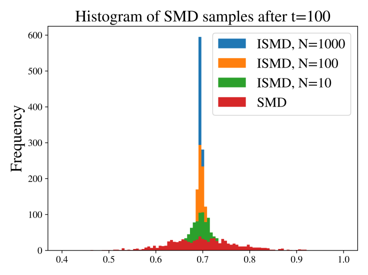

Figure: A histogram of the samples for the comparison between SMD and ISMD with 10, 100 and 1000 particles. With more particles the variance of the samples is lower.

MINI-BATCH OPTIMIZATION

- In machine learning the optimization objective typically consists of a sum over data samples, i.e. $f(x) = \sum_{i=1}^{m} f_i(x)$, where m is the sample size.
- The gradient is computed over a subset of the data (a mini-batch), since for large m computing the full gradient is too costly.
- The downside of this is that the gradient contains noise.
- Interaction is a way to decrease noise.
- But it comes at a cost! Let us compare cost of interaction against an increased batch size.

MINI-BATCH OPTIMIZATION

- Consider,

$$f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x) := \frac{1}{m}\sum_{i=1}^{m} ||W_{i,.}x - b_i||_2^2. \qquad (2)$$

- In every iteration, the gradient is computed over a subset of the data, $f_{\mathcal{S}}(x) = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} f_i(X)$, where $|\mathcal{S}|$ refers to the size of the batch.
- The optimization algorithm is then given by,

$$z_{t+1}^i = z_t^i - \eta\epsilon\nabla f_{\mathcal{S}}(x_t^i) + \epsilon\sum_{j=1}^{N} A_{ij}(z_t^j - z_t^i).$$

- The noise is thus implicit in the gradient $\nabla f_{\mathcal{S}}$.

MINI-BATCH OPTIMIZATION

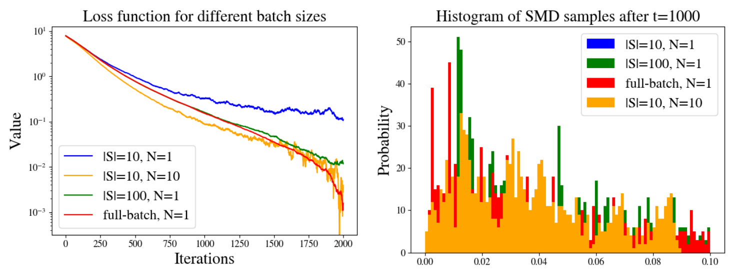

Figure: The loss function (L) and the histogram (R) for ISMD with different batch sizes with $\kappa(W) = 200$. Using interacting particles allows to use a smaller batch size while still attaining convergence. The presented results are averaged over 10 runs.

NON-CONVEX OPTIMIZATION

- Back to the setting of interest: non-convex objectives
- Consider the well-known Müller-Brown (MB)

$$f(x, y) = \sum_{i=1}^{4} A_i \exp(a_i(x - \bar{x}_i)^2 + b_i(x - \bar{x}_i)(y - \bar{y}_i) + c_i(y - \bar{y}_i)^2).$$

- It has several saddle points and local minima.
- What can interaction do?

NON-CONVEX OPTIMIZATION

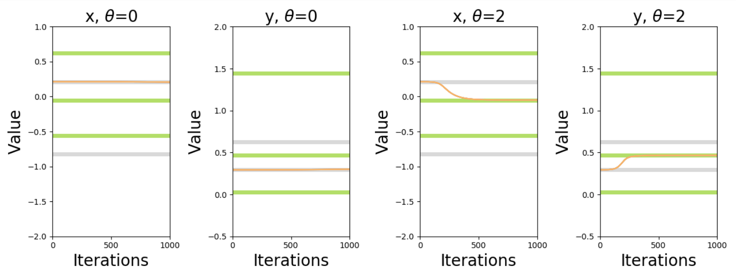

Figure: Starting from a saddle point with 10 particles using interacting SGD with a low and high interaction strength in a no-noise setting. Interactions help escape saddle points in interacting GD with learning rate $\eta = 5\text{e} - 6$.

NON-CONVEX OPTIMIZATION

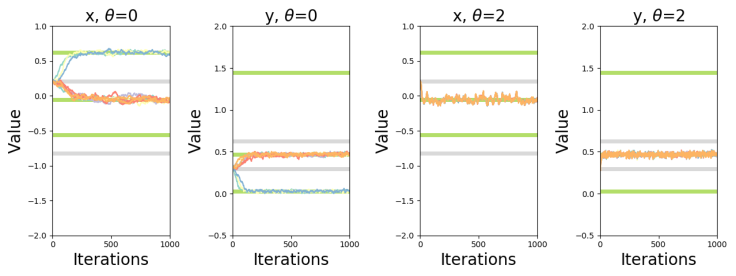

Figure: Starting from a saddle point with 10 particles using interacting SGD with a low and high interaction strength. Interaction strength imposes consensus in interacting SGD with $\sigma = 0.005$.

APPLICATIONS: FEDERATED LEARNING WITH PRIVACY-GUARANTEES

- Consider a setting in which each node in our distributed optimization setting has access to certain local data, e.g. a mobile device.
- The nodes have to communicate with each other or with a centralized server to collectively optimize the objective $f(x) = \sum_{i=1}^{N} f_i(x)$.
- To do this nodes exchange x^i or $g^i := \nabla f_i(x)$.
- Problem: both parameters as gradients can leak sensitive information about the other users.
- One solution: to each gradient that is exchanged in the system add a certain amount of noise.
- Intuition: noise adds robustness to the model; robustness can be related to differential privacy.

APPLICATIONS: OPTIMAL PERTURBATIONS TO OBFUSCATE PRIVATE DATA

- When sending data from a local device to the cloud one is also sending sensitive data.
- Ideally: reduce information content in transmitted data while conserving essential pieces of information needed for learning.

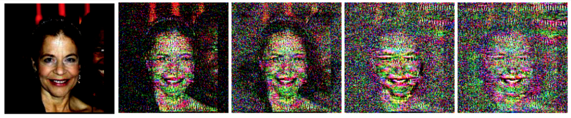

Figure: From: A Principled Approach to Learning Stochastic Representations for Privacy in Deep Neural Inference

APPLICATIONS: OPTIMAL PERTURBATIONS TO OBFUSCATE PRIVATE DATA

- This problem can be formulated as a constrained convex optimization problem:

  $$\min_{\mathcal{L}(\hat{X}) \leq x} I(X; \hat{X})$$

  where $I(X; \hat{X})$ is the mutual information between the raw input X and the data sent to the cloud $\hat{X}$ subject to the restriction on the accuracy of the prediction task x.