Tutorial on non-convex optimization with gradient methods (II):
A subjective survey of global convergence guarantees

Lénaïc Chizat*
Nov. 28nd 2019 - SMILE - Paris

*CNRS and Université Paris-Sud
The noisy way

Overdamped Langevin for optimization

Some structured problems

Geodesically convex problems

Oja’s iterations for PCA

Global convergence for neural networks

Quadratic models

Over-parameterized models with homogeneity

Lazy training
The noisy way
Gradient Langevin Dynamics

Let $f : \mathbb{R}^p \to \mathbb{R}$ smooth and satisfying $\langle \nabla f(x), x \rangle \gtrsim \|x\|^2$. Solve

$$\min_{x \in \mathbb{R}^p} f(x)$$

Discrete Overdamped Langevin Dynamics

Choose stepsize $\eta$, inverse temperature $\beta > 0$, initialisation $x_0 \in \mathbb{R}^p$. For $k \geq 1$,

1. draw $\epsilon_k \sim \mathcal{N}(0, \text{Id})$
2. $x_{t+1} = x_t - \eta \nabla f(x_t) + \sqrt{2\eta/\beta} \epsilon_k$
Guarantee

Proposition (Langevin for optimization)

There exists $C > 0$, such that with $x_0 = 0$, it holds for $k \geq 1$

$$
\mathbb{E}[f(x_k)] - \inf f \lesssim_{\beta, \eta, t} \beta^2 \exp((C - \lambda k)\eta) + \frac{\eta \log(\beta)}{\beta}
$$

where $\lambda = O(e^{-p} / \log \beta)$ is $\approx$ the Poincaré constant of the Gibbs distribution $\propto \exp(-\beta f(x)) \, dx$.

- very slow to escape stable local minima for $\beta \gg 1$
- amounts to random sampling for $\beta \ll 1$
- is noise the only hope for global continuous optimization?

[Refs]
Some structured problems
• the notion of convexity depends on the notion of straight line
• the notion of straight line depends on the metric
• some non-convex problems are convex in a different geometry
Geodesic convexity

- the notion of convexity depends on the notion of straight line
- the notion of straight line depends on the metric
- some non-convex problems are convex in a different geometry

**Definition (Geodesic convexity)**

Let $M$ a smooth, connected, $p$-dimensional Riemannian manifold.
- A closed set $\mathcal{X} \subset M$ is called $g$-convex if any two points of $M$ are joined by a unique minimizing geodesic lying in $\mathcal{X}$.
- A real-valued function $f$ on such a $\mathcal{X}$ is called $g$-convex if it is convex along constant speed minimizing geodesics.

NB: Weaker definitions exist

**Optimization problem:**

$$\min_{x \in \mathcal{X}} f(x)$$
**Algorithm: Riemannian gradient descent**

Initialize $x_0 \in \mathcal{X}$, then for $k \geq 1$ (potentially add a projection step):

$$x_{k+1} = \text{Exp}_{x_k}(-\eta_k \nabla f(x_k))$$

- enjoys similar guarantees than in the “classically convex” case
- curvature appears in the bounds (the closer to 0 the better)
- more generally: replace exponential map with first order approximations (retractions)

[Refs]
Absil, Mahony, Sepulchre (2009). *Optimization algorithms on matrix manifolds.*
Geometric programming

Consider *posynomial* functions $f_i(x) = \sum_{k=1}^{K} c_k x_1^{a_{1k}} \cdots x_p^{a_{pk}}$ where $a_{ik} \in \mathbb{R}$ and $c_k > 0$. Geometric programs are of the form

$$\min_{x \in \mathbb{R}_+^p} f_0(x) \quad \text{s.t.} \quad f_i(x) \leq 1 \ (\text{or} = 1), \quad i \in \{1, \ldots, m\}$$

- $(u_1, \ldots, u_p) \mapsto \log f_i(e^{u_1}, \ldots, e^{u_p})$ is convex
- i.e. $\log f_i$ is g-convex in the geometry induced by $x \mapsto \log(x)$
- beyond change of variable: optimization over the cone of positive semi-definite matrices with its Riemannian structure

See also: g-convex functions in Wasserstein space for sampling

[Refs]
Sra, Hosseini (2014). *Conic geometric optimisation on the manifold of positive definite matrices.*
Wibisono (2018). *Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem.*
• in these problems, convexity is hidden but present
• what can be said about really non-convex problems?
**Goal:** 1-PCA. Given a psd matrix $A \in \mathbb{R}^{p \times p}$, find $x^*$ solving

$$
\max_{\|x\|\leq 1} \frac{1}{2} x^T A x
$$

Projected gradient ascent.

Let $x_0 \sim U(S_p - 1)$. For $k \geq 1$,

$$
\begin{align*}
    y_{k+1} &= x_k + \eta_k A x_k = (I + \eta_k A)x_k \\
    x_{k+1} &= \frac{y_{k+1}}{\|y_{k+1}\|}
\end{align*}
$$

• batch version of Oja's iterations
• update of $y$ is linear: projections can be postponed

Iterates in $\mathbb{R}^2$

Iterates in $S^1$
1-PCA

**Goal:** 1-PCA. Given a psd matrix $A \in \mathbb{R}^{p \times p}$, find $x^*$ solving

$$\max_{\|x\| \leq 1} \frac{1}{2} x^T A x$$

**Projected gradient ascent.** Let $x_0 \sim \mathcal{U}(S^{p-1})$. For $k \geq 1$,

$$\begin{align*}
y_{k+1} &= x_k + \eta_k A x_k = (I + \eta_k A)x_k \\
x_{k+1} &= y_{k+1} / \|y_{k+1}\|
\end{align*}$$

- batch version of Oja’s iterations
- update of $y$ is linear: projections can be postponed

Iterates in $\mathbb{R}^2$

Iterates in $S^1$
Guarantees for the online case

Goal: online 1-PCA. Same, but given \((A_k)_{k \geq 1}\) i.i.d. \(\mathbb{E}[A_k] = A\).

Algorithm: Oja’s iterations. \(x_0 \sim \mathcal{U}(S^{p-1})\) and

\[
\begin{align*}
y_{k+1} &= x_k + \eta_k A_k x_k \\
x_{k+1} &= y_{k+1} / \|y_{k+1}\|
\end{align*}
\]
Guarantees for the online case

Goal: online 1-PCA. Same, but given \( (A_k)_{k \geq 1} \) i.i.d. \( \mathbb{E}[A_k] = A \).

Algorithm: Oja’s iterations. \( x_0 \sim \mathcal{U}(S^{p-1}) \) and

\[
\begin{align*}
  y_{k+1} &= x_k + \eta_k A_k x_k \\
  x_{k+1} &= y_{k+1} / \|y_{k+1}\|
\end{align*}
\]

Proposition (Jain et al. 2016)

Assume \( A_k \) and its moments are bounded. For a specific \( (\eta_k)_k \) decaying in \( O(1/k) \), it holds with probability \( \geq 3/4 \),

\[
\sin(x_k, x^*)^2 \lesssim \frac{\log d}{k \cdot (\lambda_1(A) - \lambda_2(A))^2} + o\left(\frac{1}{k}\right).
\]

- matches the best known statistical estimation rate
- extends to \( k \)-PCA (with orthonormalization at each iteration)

[Refs]
Allen-Zhu, Li (2017). First Efficient Convergence for Streaming \( k \)-PCA [...].
Global convergence for neural networks
Supervised machine learning

- given input/output training data \((x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\)
- build a function \(f\) such that \(f(x) \approx y\) for unseen data \((x, y)\)

Gradient-based learning

- choose a parametric class of functions \(f(w, \cdot) : x \mapsto f(w, x)\)
- a loss \(\ell\) to compare outputs: squared, logistic, cross-entropy...
- starting from some \(w_0\), update parameters using gradients

Example: Stochastic Gradient Descent with step-sizes \((\eta^{(k)})_{k \geq 1}\)

\[
w^{(k)} = w^{(k-1)} - \eta^{(k)} \nabla_w [\ell(f(w^{(k-1)}, x^{(k)}), y^{(k)})]
\]

[Refs]:
Corresponding optimization problems

We assume that \((x_i, y_i)^{i.i.d.} \sim (X, Y)\).

**Population risk.** Let \(h(w) = f(w, \cdot) \in L^2(X)\) and solve

\[
\min_w R(h(w))
\]

where \(R(f) = \mathbb{E}_{(X, Y)}[\ell(f(X), Y)]\) for \(f \in L^2(X)\).
Corresponding optimization problems

We assume that \((x_i, y_i) \overset{i.i.d.}{\sim} (X, Y)\).

**Population risk.** Let \(h(w) = f(w, \cdot) \in L^2(X)\) and solve

\[
\min_w R(h(w))
\]

where \(R(f) = \mathbb{E}_{(X, Y)}[\ell(f(X), Y)]\) for \(f \in L^2(X)\).

**Regularized empirical risk.** Let \(h(w) = f(w, x_i))_{i \in [n]}\) and solve

\[
\min_w R(h(w)) + G(w)
\]

where \(R(\hat{y}) = \frac{1}{n} \sum_{i \in [n]} \ell(\hat{y}_i, y_i)\) and \(G\) is a (convex) regularizer.

→ in the following, we write \(R\) for any convex loss on a Hilbert space.
Linear in the parameters

Linear regression, prior/random features, kernel methods:

\[ f(w, x) = w \cdot \phi(x) \]

- convex optimization
- generalization: upsides and downsides
Models

**Linear in the parameters**

Linear regression, prior/random features, kernel methods:

\[ f(w, x) = w \cdot \phi(x) \]

- convex optimization
- generalization: upsides and downsides

**Neural networks**

Vanilla NN with activation \( \sigma \) & parameters \((W_1, b_1), \ldots, (W_L, b_L)\):

\[ f(w, x) = W_L^T \sigma(W_{L-1}^T \sigma(\ldots \sigma(W_1^T x + b_1)\ldots) + b_{L-1}) + b_L \]

- interacting & interchangeable units/filters
- compositional
Consider a 2-layer neural network:

- Input $x \in \mathbb{R}^d$, output $y \in \mathbb{R}$
- Hidden weights $W = (w_1, \ldots, w_m) \in \mathbb{R}^{d \times m}$
- Fixed output layer with weights 1
- Quadratic activation function/non-linearity

$$f(W, x) = \sum_{i=1}^{m} (w_i^T x)^2$$
$$= x^T WW^T x$$
Non-convex & convex formulations

With (optionally) a regularization, we obtain a problem of the form

$$\min_{W \in \mathbb{R}^{d \times m}} R(WW^T) + \lambda \|W\|_F^2$$  \hspace{1cm} (1)

where $R$ is a convex function and $\lambda \geq 0$.

If $m \geq d$, posing $M = WW^T$, equivalent to the convex problem

$$\min_{M \in \mathbb{S}^{d \times d}_+} R(M) + \lambda \|M\|_*$ \hspace{1cm} (2)$$
Lemma

Any rank deficient 2\textsuperscript{nd}-order stationary point of (1) is a minimizer.

Theorem (Nice landscape)

If \( m \geq d \), 2\textsuperscript{nd}-order stationary points of (1) are minimizers.

Proof.

If \( \text{rk}(W) < d \) then the lemma applies. If \( \text{rk}(W) = d \), then 1st order optimality for (1) implies first order optimality for (2).

[Refs]
Haeffele, Young, Vidal (2014). Structured Low-Rank Matrix Factorization [...]
Remarks

- related to the guarantees on Burer-Monteiro factorization for low-rank semi-definite programming
- instance of a landscape analysis approach
- next step: analysis of optimization paths

[Refs]
Two-layer neural network

With activation $\sigma$, define $\phi(w_i, x) = c_i \sigma(a_i \cdot x + b_i)$ and

$$f(w, x) = \frac{1}{m} \sum_{i=1}^{m} \phi(w_i, x)$$

**Difficulty**: existence of spurious minima, e.g. for the population loss even with slight over-parameterization

[Refs]:
Livni, Shalev-Shwartz, Shamir (2014). *On the Computational Efficiency of Training Neural Networks.*
Safran, Shamir (2018). *Spurious Local Minima are Common in Two-layer ReLU Neural Networks.*
• let $\mathcal{P}_2(\mathbb{R}^p)$ be the space of probability measures endowed with the 2-Wasserstein metric.

• consider the model

$$f(\mu, x) = \int \phi(w, x) \, d\mu(w).$$

• as before, let $h(\mu) = f(\mu, \cdot)$ and solve

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^p)} F(\mu) \quad \text{where} \quad F(\mu) = R(h(\mu))$$

• regularization possible but skipped for simplicity

[Refs]:
Many-particle / mean-field limit

Measure representation

The gradient flow \((w(t))_{t \geq 0}\) on the objective defines a dynamics in the space \(\mathcal{P}_2(\mathbb{R}^p)\) of probabilities endowed with the Wasserstein metric:

\[
\mu_{t,m} = \frac{1}{m} \sum_{i=1}^{m} \delta_{w_i(t)}
\]

Theorem (Many particle limit)

Assume that \(w_1(0), w_2(0), \ldots\) are such that \(\mu_{0,m} \to \mu_0\) and technical assumptions. Then \(\mu_{t,m} \to \mu_t\), uniformly on \([0, T]\), where \(\mu_t\) is the unique Wasserstein gradient flow of \(F\) starting from \(\mu_0\).

[Refs]:
Rotskoff, Vanden-Eijndem (2018). Parameters as Interacting Particles [...].
Sirignano, Spiliopoulos (2018). Mean Field Analysis of Neural Networks.

Global convergence

**Theorem (2-homogeneous case)**

Assume that $\phi$ is positively 2-homogeneous and technical assumptions. If the support of $\mu_0$ covers all directions (e.g. Gaussian) and if $\mu_t \to \mu_\infty$, then $\mu_\infty$ is a global minimizer of $F$.

$\leadsto$ Non-convex landscape : initialization matters
Global convergence

**Theorem (2-homogeneous case)**

Assume that $\phi$ is positively 2-homogeneous and technical assumptions. If the support of $\mu_0$ covers all directions (e.g. Gaussian) and if $\mu_t \to \mu_\infty$, then $\mu_\infty$ is a global minimizer of $F$.

$\leadsto$ Non-convex landscape : initialization matters

**Corollary**

Under the same assumptions, if at initialization $\mu_{0,m} \to \mu_0$ then

$$\lim_{t \to \infty} \lim_{m \to \infty} F(\mu_{m,t}) = \lim_{m \to \infty} \lim_{t \to \infty} F(\mu_{m,t}) = \inf F.$$  

[Refs]:
ReLU, $d = 2$, optimal predictor has 5 neurons (population risk)
Infinite width limit of standard neural networks

For infinitely wide fully connected neural networks of any depth with “standard” initialization and no regularization: the gradient flow implicitly performs kernel ridge(less) regression with the neural tangent kernel

\[ K(x, x') = \lim_{m \to \infty} \langle \nabla_w \tilde{f}_m(w_0, x), \nabla_w \tilde{f}_m(w_0, x') \rangle. \]
Neural Tangent Kernel (Jacot et al. 2018)

Infinite width limit of standard neural networks

For infinitely wide fully connected neural networks of any depth with “standard” initialization and no regularization: the gradient flow implicitly performs kernel ridge(less) regression with the neural tangent kernel

\[ K(x, x') = \lim_{m \to \infty} \langle \nabla_w \tilde{f}_m(w_0, x), \nabla_w \tilde{f}_m(w_0, x') \rangle. \]

Reconciling the two views:

\[ \tilde{f}_m(w, x) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \phi(w_i, x) \quad vs. \quad f_m(w, x) = \frac{1}{m} \sum_{i=1}^{m} \phi(w_i, x) \]

This behavior is not intrinsically due to over-parameterization but to an exploding scale

[Refs]:
Linearized model and scale

- Let $h(w) = f(w, \cdot)$ be a differentiable model.
- Let $\bar{h}(w) = h(w_0) + Dh_{w_0}(w - w_0)$ be its linearization at $w_0$.

Compare 2 training trajectories starting from $w_0$, with scale $\alpha > 0$:

- $w_\alpha(t)$ gradient flow of $F_\alpha(w) = R(\alpha h(w))/\alpha^2$
- $\bar{w}_\alpha(t)$ gradient flow of $\bar{F}_\alpha(w) = R(\alpha \bar{h}(w))/\alpha^2$

$\Rightarrow$ if $h(w_0) \approx 0$ and $\alpha$ large, then $w_\alpha(t) \approx \bar{w}_\alpha(t)$
Lazy training theorems

**Theorem (Non-asymptotic)**

*If \( h(w_0) = 0 \), and \( R \) potentially non-convex, for any \( T > 0 \), it holds*

\[
\lim_{\alpha \to \infty} \sup_{t \in [0, T]} \| \alpha h(w_\alpha(t)) - \alpha \tilde{h}(\tilde{w}_\alpha(t)) \| = 0
\]

**Theorem (Strongly convex)**

*If \( h(w_0) = 0 \), and \( R \) strongly convex, it holds*

\[
\lim_{\alpha \to \infty} \sup_{t \geq 0} \| \alpha h(w_\alpha(t)) - \alpha \tilde{h}(\tilde{w}_\alpha(t)) \| = 0
\]

- instance of *implicit bias*: lazy because parameters hardly move
- may replace the model by its linearization
- commonly used neural networks are not in this regime

[Refs]:
Conclusion
Conclusion

Not covered

- restricted convexity/isometry property for sparse regression
- other methods: alternating minimization, relaxations,...

Opening remarks

- in the non-convex world, no unifying theory but a variety of structures coming with a variety of guarantees
- renewed theoretical interest due to good practical behavior and good scalability (compared e.g. to convex relaxations)

[Refs]