Nonconvex Global Optimization

Technische Universität München

January 6, 2025

Contents

1 Optimization of smooth and convex functions by gradient descent

We consider a function $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}$ (the objective function), $\mathcal{E} \in C^2(\mathbb{R}^d)$, with Lipschitz continuous gradient, i.e., there exists a $L > 0$ such that

$$\|\nabla\mathcal{E}(x) - \nabla\mathcal{E}(y)\| \leq L\|x - y\| \quad \forall x, y \in \mathbb{R}^d. \tag{1.1}$$

We wish to minimize $\mathcal{E}$, i.e., we want to find

$$x^* \in \arg\min_{x\in\mathbb{R}^d} \mathcal{E}(x). \tag{1.2}$$

In order to find such a minimizer we also consider the time-dependent process

$$\begin{cases} \dot{x}(t) = \frac{d}{dt}x(t) = -\nabla\mathcal{E}(x(t)) \\ x(0) = x_0 \end{cases} \tag{1.3}$$

Starting at some point x_0, we can move along the direction of greatest descent $-\nabla\mathcal{E}$ until after some time $\bar{t} > 0$ we come to a point with $\dot{x}(\bar{t}) = 0$ and the process becomes stationary. For the point $\bar{x} = x(\bar{t})$, we have $\nabla\mathcal{E}(\bar{x}) = 0$, i.e., it is a critical point of $\mathcal{E}$, and therefore it might be a (local) minimizer of $\mathcal{E}$.

The integral formulation of the Ordinary Differential Equation (ODE) (1.3) reads

$$x(t) = x_0 + \int_0^t -\nabla\mathcal{E}(x(s))\, ds = T_{x_0}(x)(t),$$

which can be written in a fixed point formulation

$$x = T_{x_0}(x).$$

By the Picard-Lindelöf iteration (or Banach's fixed point theorem), one can show that the ODE (1.3) possesses a unique solution $t \mapsto x(t)$.

Lemma 1.1. *For a smooth function $\mathcal{E}:\mathbb{R}^d\to\mathbb{R}$ with L-Lipschitz continuous gradient it holds*

$$|\mathcal{E}(y)-\mathcal{E}(x)-\langle\nabla\mathcal{E}(x),y-x\rangle|\le\frac{L}{2}\|x-y\|^2\quad\forall x,y\in\mathbb{R}^d. \tag{1.4}$$

Remark. *From Lemma 1.1 it immediately follows*

$$\mathcal{E}(y)\le\mathcal{E}(x)+\langle\nabla\mathcal{E}(x),y-x\rangle+\frac{L}{2}\|x-y\|^2. \tag{1.5}$$

Lemma 1.2. *Let $\mathcal{E}:\mathbb{R}^d\to\mathbb{R}$ be smooth with a L-Lipschitz continuous gradient. Further assume that there exists a global minimizer x^* of $\mathcal{E}$. Then the inequalities*

$$\mathcal{E}(x^*)\le\mathcal{E}\left(x-\frac{1}{L}\nabla\mathcal{E}(x)\right)\le\mathcal{E}(x)-\frac{1}{2L}\|\nabla\mathcal{E}(x)\|^2$$

are true for every $x\in\mathbb{R}^d$.

Lemma 1.3 (Co-Coercivity of the gradient)**.** *For a smooth and convex function $\mathcal{E}:\mathbb{R}^d\to\mathbb{R}$ with L-Lipschitz continuous gradient it holds*

$$\|\nabla\mathcal{E}(x)-\nabla\mathcal{E}(y)\|^2\le L\langle x-y,\nabla\mathcal{E}(x)-\nabla\mathcal{E}(y)\rangle\quad\forall x,y\in\mathbb{R}^d. \tag{1.7}$$

Remark. *From Lemma 1.3 it follows:*

- $\langle x-y,\nabla\mathcal{E}(x)-\nabla\mathcal{E}(y)\rangle\ge0$ *which is equivalent to the convexity of $\mathcal{E}$ (see Lemma A.1).*
- *By Cauchy-Schwarz inequality we get*

$$\langle x-y,\nabla\mathcal{E}(x)-\nabla\mathcal{E}(y)\rangle\le\|x-y\|\|\nabla\mathcal{E}(x)-\nabla\mathcal{E}(y)\|.$$

Therefore co-coercivity implies Lipschitz continuity of $\nabla\mathcal{E}$.

Now we show that under the assumption of strong convexity the trajectory $t\mapsto x(t)$ of gradient descent (1.3) *converges to a minimizer of* (1.2)*.*

Theorem 1.4 (Convergence of gradient descent)**.** *Assume $\mathcal{E}\in C^2(\mathbb{R}^d)$ is γ-strongly convex, i.e., there exists a $\gamma>0$ such that*

$$\nabla^2\mathcal{E}(x)\succeq\gamma I,$$

and that there exists some $L\ge\gamma>0$ such that

$$\nabla^2\mathcal{E}(x)\preceq LI.$$

Then the evolution $t\mapsto x(t)$ solution of the ODE

$$\begin{cases}\dot{x}(t)=\frac{d}{dt}x(t)=-\nabla\mathcal{E}(x(t))\\ x(0)=x_0\end{cases}$$

always converges to the unique minimizer

$$x^*\in\arg\min_{x\in\mathbb{R}^d}\mathcal{E}(x).$$

Remark. *The condition $\nabla^2\mathcal{E}(x)\preceq LI$ in Theorem 1.4 is equivalent to $\nabla\mathcal{E}$ being L-Lipschitz continuous. By the Picard-Lindelöf theorem this implies that there exists a unique solution $x(t)$ of the ODE which is implicitly assumed in Theorem 1.4.*

For the speed of convergence we consider again

$$\begin{aligned}\frac{d}{dt}f(x(t))&=-\langle x(t)-x^*,\nabla\mathcal{E}(x(t))-\nabla\mathcal{E}(x^*)\rangle\\&\le-\gamma\|x(t)-x^*\|^2\\&=-2\gamma f(x(t)),\end{aligned}$$

where we used the strong convexity in the first inequality (see Lemma A.1). Applying Gronwall's inequality gives

$$\|x(t)-x^*\|^2\le\|x_0-x^*\|^2e^{-2\gamma t}\to0\quad\text{exponentially fast.}$$

So far we have never tried $\mathcal{E}$ itself as a Lyapunov function. So let us try it:

$$\frac{d}{dt}\mathcal{E}(x(t))=-\left\langle\nabla\mathcal{E}(x(t)),\frac{d}{dt}x(t)\right\rangle$$

$$= \langle \nabla \mathcal{E}(x(t)), -\nabla \mathcal{E}(x(t)) \rangle$$
$$= -\|\nabla \mathcal{E}(x(t))\|^2 \leq 0.$$

Let us assume that there exists some $\kappa > 0$ such that $\mathcal{E}$ fulfills

$$\|\nabla \mathcal{E}(x)\|^2 \geq 2\kappa(\mathcal{E}(x) - \mathcal{E}(x^*))$$

for all x (this is called the Polyak-Łojasiewicz inequality which we consider in more detail in Section 4). Then

$$\frac{d}{dt}(\mathcal{E}(x(t)) - \mathcal{E}(x^*)) = \frac{d}{dt}\mathcal{E}(x(t))$$
$$= -\|\nabla \mathcal{E}(x(t))\|^2$$
$$\leq -2\kappa(\mathcal{E}(x(t)) - \mathcal{E}(x^*)).$$

and by applying Gronwall's inequality we get

$$\mathcal{E}(x(t)) - \mathcal{E}(x^*) \leq (\mathcal{E}(x_0) - \mathcal{E}(x^*))e^{-2\kappa t}$$

and therefore also exponential convergence.

Note that the Polyak-Łojasiewicz inequality does not imply convexity of $\mathcal{E}$ (a counterexample is given in Example 4.4). However it does imply that $\mathcal{E}$ does not have spurious local minima. Therefore gradient descent also converges for some non-convex functions.

2 Gradient descent algorithm

Formally gradient descent algorithm is an Euler discretization (in time) of

$$\dot{x} = -\nabla \mathcal{E}(x)$$

and reads as follows:

$$\begin{cases} x^{(n+1)} = x^{(n)} - \alpha^{(n)} \nabla \mathcal{E}(x^{(n)}), & n \geq 0 \\ x^{(0)} = x_0 \end{cases} \tag{GD}$$

In fact, one can rewrite

$$x^{(n)} = \frac{d}{dt}x(t_n) \approx \frac{x^{(n+1)} - x^{(n)}}{\alpha^{(n)}} \quad \text{(finite difference approximation of a time derivative)}$$
$$= -\nabla \mathcal{E}(x^{(n)}).$$

It is also an explicit approximation of the implicit minimizing movement scheme

$$x^{(n+1)} = \arg\min_{x \in \mathbb{R}^d} \left\{ \frac{\|x - x^{(n)}\|^2}{2\alpha^{(n)}} + \mathcal{E}(x) \right\},$$

which is equivalent to the implicit Euler scheme since $x^{(n+1)}$ has to fulfill the necessary first-order condition

$$\frac{d}{dx}\left[\frac{\|x - x^{(n)}\|^2}{2\alpha^{(n)}} + \mathcal{E}(x) \right] = 0$$
$$\iff \frac{x - x^{(n)}}{\alpha^{(n)}} + \nabla \mathcal{E}(x) = 0,$$

which leads to

$$x^{(n+1)} = x^{(n)} - \alpha^{(n)} \nabla \mathcal{E}(x^{(n+1)}).$$

How do we compare $x(t_n)$ with $x^{(n)}$ from (GD)? In general and without further information we can only expect (by numerical analysis)

$$\|x(t_n) - x^{(n)}\| \leq C^n \max_{1 \leq k \leq n} \alpha^{(k)}$$

for some constant $C > 0$.

Theorem 2.1 (Convergence of (GD))**.** *Assume $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}$ is smooth and convex with L-Lipschitz continuous gradient. Further assume that there exists a (global) minimizer x^* of $\mathcal{E}$. Then the function values of the iterates $x^{(n)}$ of (GD) converge for any sequence $(\alpha^{(n)})_{n \in \mathbb{N}}$ with $\tau < \alpha^{(n)} < \frac{1}{L}$ for all $n \geq 0$ and some $\tau > 0$ to the optimal value, i.e.,*

$$\lim_{n \to \infty} \mathcal{E}(x^{(n)}) = \mathcal{E}(x^*). \tag{2.1}$$

Proof. □

Remark.

- ***Theorem 2.1*** *does not imply that* $x(t_n) \approx x^{(n)}$ *holds for all* $n \geq 0$.
- *In the case of multiple global minimizers we can only expect convergence of the function values of the iterates but not convergence of the iterates themselves.*

3 Stochastic gradient descent algorithm

We assume that $(x_j, y_j) \overset{\text{iid}}{\sim} P, j = 1, \dots, N$ for some probability distribution P and consider the misfit over all given examples

$$\frac{1}{N}\sum_{j=1}^{N} f(\theta, (x_j, y_j)) := \mathcal{E}_N(\theta) \quad \text{with} \quad \omega_j = (x_j, y_j),$$

where $f(\theta, \omega_j)$ measures the misfit (for example one such measure could be $f(\theta, \omega_j) = |\mathcal{A}(x_j, \theta) - y_j|^2$). If we now have $N \to \infty$ examples then we hope that

$$\mathcal{E}_N(\theta) \to \int_\Omega f(\theta, \omega)\, dP(\omega) := \mathcal{E}(\theta).$$

$\mathcal{E}(\theta)$ can be considered as the misfit of the model $(\mathcal{A}(x, \theta), y)$ integrated over all the statistics. The final goal is always to find a θ^* that makes $\mathcal{E}(\theta)$ the smallest possible. Since we do not know all statistics one approach could be to compute $\theta_N^* = \arg\min_\theta \mathcal{E}_N(\theta)$ and use θ_N^* as a possible approximate solution to $\arg\min_\theta \mathcal{E}(\theta)$.

Assume we have a random variable $f(\theta, \omega)$ where $\omega \in (\Omega, \mathcal{F}, P)$ and consider

$$\mathcal{E}(\theta) = \mathbb{E}[f(\theta, \omega)] = \int_\Omega f(\theta, \omega)\, dP(\omega).$$

Minimizing this function by gradient descent requires being able to compute

$$\nabla_\theta \mathcal{E}(\theta) = \nabla_\theta \int_\Omega f(\theta, \omega)\, dP(\omega) = \int_\Omega \nabla_\theta f(\theta, \omega)\, dP(\omega)$$

(under sufficient regularity assumptions). This requires being able to access $\nabla_\theta f(\theta, \omega)$ for every $\omega \in \Omega$. We may approximate:

- If we have enough examples then we can approximate

$$\nabla_\theta \mathcal{E}(\theta) \approx \frac{1}{N}\sum_{j=1}^{N} \nabla_\theta f(\theta, \omega_j),$$

  with $\omega_j \overset{\text{iid}}{\sim} P, j = 1, \dots, N$.
- If $\theta^{(1)}, \dots, \theta^{(N)} \approx \theta$, then

$$\nabla_\theta \mathcal{E}(\theta) \approx \frac{1}{N}\sum_{j=1}^{N} \nabla_\theta f(\theta^{(j)}, \omega_j).$$

These approximations suggest the following empirical instances of (GD), the so-called **Stochastic Gradient Descent**:

$$\begin{cases} \theta^{(n+1)} = \theta^{(n)} - \alpha^{(n)} \nabla_\theta f\left(\theta^{(n)}, \omega_n\right), \quad n \geq 0, \\ \theta^{(0)} = \theta_0, \\ \omega_n \overset{\text{iid}}{\sim} P \end{cases} \tag{SGD}$$

While we are iterating we get:

$$\begin{aligned} \theta^{(n+1)} &= \theta^{(n)} - \alpha^{(n)} \nabla_\theta f\left(\theta^{(n)}, \omega_n\right) \\ &= \theta^{(n-1)} - \alpha^{(n-1)} \nabla_\theta f\left(\theta^{(n-1)}, \omega_{n-1}\right) - \alpha^{(n)} \nabla_\theta f\left(\theta^{(n)}, \omega_n\right) \\ &= \theta^{(n-N)} - \underbrace{\sum_{j=0}^{N} \alpha^{(n-j)} \nabla_\theta f\left(\theta^{(n-j)}, \omega_{n-j}\right)}_{\approx \alpha \nabla_\theta \mathcal{E}(\theta)} \\ &\approx \nabla_\theta \mathcal{E}(\theta). \end{aligned}$$

We hope that the expected behavior of the algorithm can in turn be the minimization of $\mathcal{E}$. Now let us make our ideas more precise and prove them in the strongly convex regime.

Theorem 3.1 (Convergence of SGD). *Let each of $\theta \mapsto f(\theta, \omega)$ be convex with $\nabla_\theta f(\cdot, \omega)$ being $L(\omega)$-Lipschitz continuous with*

$$L := \sup_{\omega \in \Omega} L(\omega) < +\infty.$$

Let also

$$\mathcal{E}(\theta) = \mathbb{E}_\omega \left[f(\theta, \omega) \right] = \int_\Omega f(\theta, \omega)\, dP(\omega)$$

be γ-strongly convex (note that this implies $\gamma < L$). Set

$$\sigma^2 = \mathbb{V} \left[\nabla_\theta f(\theta^*, \omega) \right] = \mathbb{E} \left[\left(\nabla_\theta f(\theta^*, \omega) \right)^2 \right]$$

*for $\theta^* = \arg\min_{\theta \in \mathbb{R}^d} \mathcal{E}(\theta)$. For simplicity assume $\alpha^{(n)} = \alpha < \frac{1}{L}$. Then the iterates of (**SGD**) satisfy the following estimate:*

$$\mathbb{E} \left[\|\theta^{(n)} - \theta^*\|_2^2 \right] \leq (1 - 2\alpha\gamma(1 - \alpha L))^n \, \|\theta^{(0)} - \theta^*\|_2^2 + \frac{\alpha \sigma^2}{\gamma(1 - \alpha L)} \tag{3.2}$$

where the expected value is over $\omega^{(0)}, \ldots, \omega^{(n)}, \ldots \overset{iid}{\sim} P$.

Remark. *Because of $\alpha < \frac{1}{L}$, it holds*

$$1 - 2\alpha\gamma \underbrace{(1 - \alpha L)}_{> o} < 1$$

and because of $\gamma < L$, it holds

$$1 - 2\alpha\gamma(1 - \alpha L) = 1 - 2\alpha\gamma + 2\alpha^2\gamma L \leq 1 - 2\alpha\gamma + 2\alpha^2\gamma^2 = (\alpha\gamma - 1)^2 + \alpha^2\gamma^2 > 0$$

which results in

$$0 < 1 - 2\alpha\gamma(1 - \alpha L) < 1.$$

Suppose we know the problem-specific constants γ, L, θ^*, and σ^2, then we can search for good hyperparameters α and n to obtain convergence

$$\mathbb{E} \left[\|\theta^{(n)} - \theta^*\|_2^2 \right] \leq \epsilon$$

for any given tolerance $\epsilon > 0$.

Corollary 3.2. *For any $\epsilon > 0$, fix*

$$\alpha^* = \frac{\epsilon\gamma}{2\epsilon\gamma L + 2\sigma^2}.$$

Then after

$$n \geq 2 \log \left(\frac{2\epsilon_0}{\epsilon} \right) \left(\frac{L}{\gamma} + \frac{\sigma^2}{\gamma^2 \epsilon} \right)$$

*iterations, (**SGD**) fulfills*

$$\mathbb{E} \left[\|\theta^{(n)} - \theta^*\|_2^2 \right] \leq \epsilon$$

where we set

$$\epsilon_0 = \|\theta^{(0)} - \theta^*\|_2^2.$$

4 Invex objective functions

Definition 4.1 (invex function). *A function $f : \mathbb{R}^n \to \mathbb{R}$ is called **invex** if there exists a vector-valued function $\eta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:*

$$f(\mathbf{y}) - f(\mathbf{x}) \geq \nabla f(\mathbf{x})^\top \eta(\mathbf{x}, \mathbf{y}).$$

Here:

- *$\nabla f(\mathbf{x})$ is the gradient of f at $\mathbf{x}$,*
- *$\eta(\mathbf{x}, \mathbf{y})$ is called the **invexity vector**.*

we have another kind of definition for smooth function:

Definition 4.2. *If f is a **smooth function** (i.e., continuously differentiable), then f is **invex** if and only if: Every critical point of f is a global minimizer.*

A critical point $\mathbf{x}^*$ satisfies $\nabla f(\mathbf{x}^*) = \mathbf{0}$. If every critical point is a global minimizer, the function is invex.

Proof. ($\Rightarrow$): If f is invex according to the original definition, we want to prove that every critical point of f is a global minimizer.

1. Let $\mathbf{x}^*$ be a critical point of f, so:
$$\nabla f(\mathbf{x}^*) = \mathbf{0}.$$
2. From the definition of invexity, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:
$$f(\mathbf{y}) - f(\mathbf{x}) \geq \nabla f(\mathbf{x})^\top \eta(\mathbf{x}, \mathbf{y}).$$
3. Substitute $\mathbf{x} = \mathbf{x}^*$. Since $\nabla f(\mathbf{x}^*) = \mathbf{0}$, the inequality becomes:
$$f(\mathbf{y}) - f(\mathbf{x}^*) \geq 0.$$
4. Rearranging, we get:
$$f(\mathbf{y}) \geq f(\mathbf{x}^*), \quad \forall \mathbf{y} \in \mathbb{R}^n.$$
5. This shows that $\mathbf{x}^*$ is a global minimizer.

Thus, the original definition implies that every critical point is a global minimizer.
($\Leftarrow$): If every critical point of f is a global minimizer, we want to prove that f satisfies the original invexity condition.

1. Assume $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are arbitrary points. Define:
$$\eta(\mathbf{x}, \mathbf{y}) = \mathbf{y} - \mathbf{x}.$$
2. From the second definition, we know that every critical point is a global minimizer. To prove invexity, we need:
$$f(\mathbf{y}) - f(\mathbf{x}) \geq \nabla f(\mathbf{x})^\top \eta(\mathbf{x}, \mathbf{y}).$$
3. If $\mathbf{x}$ is a critical point, then $\nabla f(\mathbf{x}) = \mathbf{0}$, so:
$$f(\mathbf{y}) - f(\mathbf{x}) \geq 0,$$
which is trivially true for any $\eta(\mathbf{x}, \mathbf{y})$.
4. If $\mathbf{x}$ is not a critical point, consider the function:
$$g(t) = f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \quad t \in [0, 1].$$
   - $g(t)$ represents the value of f along the line segment between $\mathbf{x}$ and $\mathbf{y}$.
   - The derivative of $g(t)$ at $t = 0$ is:
$$g'(0) = \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}).$$
5. Using the Mean Value Theorem, there exists $t^* \in (0, 1)$ such that:
$$f(\mathbf{y}) - f(\mathbf{x}) = g'(t^*).$$
6. If every critical point is a global minimizer, $g'(t^*) \geq g'(0)$, which implies:
$$f(\mathbf{y}) - f(\mathbf{x}) \geq \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}).$$

Thus, the second definition implies the original definition. □

Definition 4.3 (Polyak-Łojasiewicz Inequality)**.** *let $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}$ be a smooth function and assume that there exists a global minimizer x^*. then $\mathcal{E}$ satisfies the Polyak-Łojasiewicz Inequality if there exists a $\mu > 0$ s.t.*

$$\mu(\mathcal{E}(x) - \mathcal{E}(x^*)) \leq \frac{1}{2}\|\nabla \mathcal{E}(x)\|_2^2 \quad \forall x \in \mathbb{R}^d \tag{4.1}$$

the Polyak-Łojasiewicz Inequality implies that every critical point of $\mathcal{E}$ is a global minimizer of $\mathcal{E}$ and therefore such functions are invex functions. However it does not imply convexity as the following examples shows.

Example 4.4. $\mathcal{E}:\mathbb{R}\to\mathbb{R}$, $\mathcal{E}(x)=x^2+2\sin^2(x)$ *with*

$$\nabla\mathcal{E}(x)=2x+3\sin(2x),\nabla^2\mathcal{E}(x)=2+6\cos(2x)$$

and the global minimizer $x^*=0$. *then this function is not convex as* $\nabla^2\mathcal{E}(\frac{\pi}{2})=-4<0$, *but it fulfills the Polyak-Łojasiewicz Inequality with* $\mu=\frac{1}{32}$ *since we have* $\forall x\in\mathbb{R}\setminus\{0\}$: $\frac{1}{2\mu}\|\nabla E(x)\|_2^2-(\mathcal{E}(x)-\mathcal{E}(x^*))=16(2x+3\sin(2x))^2-(x^2+3\sin^2(x)-0)\geq 64x^2(1+3\frac{\sin(2x)}{2x})^2-4x^2\geq 64x^2(1-\frac{3}{4})^2-4x^2=0$, *where we used* $|\sin(x)|\leq|x|$ *in the first equality and* $\frac{\sin(x)}{x}>-\frac{1}{4},\forall x\in\mathbb{R}$

Lemma 4.5 (Gronwall's inequality)**.** *Let* $u:[0,\infty)\to\mathbb{R}$ *be a differentiable function and* $\kappa:[0,\infty)\to\mathbb{R}$ *be continuous. If* u *satisfies*

$$\frac{d}{dt}u(t)\leq\kappa(t)u(t)\quad\forall t\in(0,\infty)$$

then it holds

$$u(t)\leq u(0)\exp\left(\int_0^t\kappa(s)\,ds\right).$$

If the function κ *is constant, i.e.,* $\kappa(t)\equiv\kappa$, *then the inequality simplifies to*

$$u(t)\leq u(0)e^{\kappa t}\quad\forall t\in[0,\infty).$$

Let us consider again gradient descent

$$\begin{cases}\dot{x}(t)=\frac{d}{dt}x(t)=-\nabla\mathcal{E}(x(t))\\ x(0)=x_0\end{cases}$$

where $\mathcal{E}$ satisfies the Polyak–Łojasiewicz inequality for some $\mu>0$. Then we get

$$\frac{d}{dt}\left(\mathcal{E}(x(t))-\mathcal{E}(x^*)\right)=\nabla\mathcal{E}(x(t))\cdot\frac{d}{dt}x(t)=-\|\nabla\mathcal{E}(x(t))\|_2^2\leq-2\mu\left(\mathcal{E}(x(t))-\mathcal{E}(x^*)\right)\leq 0$$

where we used the Polyak–Łojasiewicz inequality in the last step. By Gronwall's inequality (let $\kappa=-2\mu$) we get

$$\mathcal{E}(x(t))-\mathcal{E}(x^*)\leq\left(\mathcal{E}(x_0)-\mathcal{E}(x^*)\right)e^{-2\mu t}\xrightarrow{t\to\infty}0$$

and therefore convergence of gradient descent to the set of global minimizers.

Remark. *If* $\mathcal{E}(x)$ *fulfills the Polyak–Łojasiewicz inequality for some* $\mu>0$ *and has a L-Lipschitz continuous gradient then it holds* $\mu\leq L$ *since we have for every* $x\in\mathbb{R}^d$

$$\frac{1}{2}\|\nabla\mathcal{E}(x)\|^2\geq\mu\left(\mathcal{E}(x)-\mathcal{E}(x^*)\right)\geq\mu\left(\mathcal{E}(x)-\left(\mathcal{E}(x)-\frac{1}{2L}\|\nabla\mathcal{E}(x)\|^2\right)\right)=\frac{\mu}{2L}\|\nabla\mathcal{E}(x)\|^2$$

where we used the Polyak–Łojasiewicz inequality in the first step and Lemma1.2 in the second step.

we not only have convergence of gradient descent but also convergence of the (GD) algorithm.

Theorem 4.6. *Let* $\mathcal{E}:\mathbb{R}^d\to\mathbb{R}$ *with* $\mathcal{E}\in C^1(\mathbb{R}^d)$ *and L-Lipschitz continuous gradient be given and assume that*

$$\mathcal{X}^*=\arg\min_{x\in\mathbb{R}^d}\mathcal{E}(x)\neq\emptyset$$

and that $\mathcal{E}$ *fulfills the Polyak–Łojasiewicz inequality. Then (GD) with stepsize* $\alpha=\frac{1}{L}$ *has global linear convergence*

$$\mathcal{E}(x^{(n)})-\mathcal{E}(x^*)\leq\left(1-\frac{\mu}{L}\right)^n\left(\mathcal{E}(x^{(0)})-\mathcal{E}(x^*)\right)\tag{4.2}$$

for any $x^*\in\mathcal{X}^*$.

Proof. □

Note that under the assumption of the Polyak-Łojasiewicz inequality the proofs of convergence of gradient descent and (GD) are simpler than under the stronger assumption of (strong) convexity.

4.1 Convergence of GD methods for high-dimensional problems

Recall that a Lipschitz continuous gradient implies (by Lemma 1.1)

$$\mathcal{E}(y) \leq \mathcal{E}(x) + \langle \nabla \mathcal{E}(x), y - x \rangle + \frac{L}{2}\|y - x\|^2.$$

Now let us consider a coordinate-wise version of such estimate with $y = x + \alpha e_i$

$$\mathcal{E}(x + \alpha e_i) \leq \mathcal{E}(x) + \alpha \frac{\partial}{\partial x_i}\mathcal{E}(x) + \frac{L}{2}\alpha^2$$

for $i = 1, \ldots, d$ and for any $\alpha \in \mathbb{R}$ and $x \in \mathbb{R}^d$ (e_i denotes the i-th canonical unit vector).

Theorem 4.7. *Assume that $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}$ fulfills*

$$\mathcal{E}(x + \alpha e_i) \leq \mathcal{E}(x) + \alpha \frac{\partial}{\partial x_i}\mathcal{E}(x) + \frac{L}{2}\alpha^2 \quad \forall i \in \{1, \ldots, d\}, \forall \alpha \in \mathbb{R}, \forall x \in \mathbb{R}^d \tag{4.3}$$

for some $L > 0$. Further assume that $\mathcal{X}^* = \arg\min_{x \in \mathbb{R}^d} \mathcal{E}(x) \neq \emptyset$ and that $\mathcal{E}$ fulfills the Polyak–Łojasiewicz inequality with $\mu > 0$. We consider the coordinate gradient descent algorithm with stepsize $\frac{1}{L}$

$$x^{(n+1)} = x^{(n)} - \frac{1}{L}\frac{\partial}{\partial x_{i_n}}\mathcal{E}(x^{(n)}) \cdot e_{i_n}, \tag{4.4}$$

where i_n is drawn uniformly at random in $\{1, \ldots, d\}$. Then the algorithm fulfills

$$\mathbb{E}\left[\mathcal{E}(x^{(n)}) - \mathcal{E}(x^*)\right] \leq \left(1 - \frac{\mu}{Ld}\right)^n \left(\mathcal{E}(x^{(0)}) - \mathcal{E}(x^*)\right) \tag{4.5}$$

for any $x^* \in \mathcal{X}^*$.

Proof. □

So far we have shown convergence of gradient descent and (GD) algorithm under the Polyak–Łojasiewicz inequality. What about the (SGD) algorithm? Remember that we considered

$$\mathcal{E}(x) = \mathbb{E}\left[f(x, \omega)\right] = \int_\Omega f(x, \omega)\, \mathrm{d}\mathbb{P}(\omega).$$

To make it slightly simpler and more intuitive we consider the case of a discrete uniform distribution on $\Omega = \{1, \ldots, k\}$. In this case we have

$$\mathcal{E}(x) = \frac{1}{k}\sum_{i=1}^{k} f_i(x)$$

with $f_i(x) = f(x, \omega = i)$. The (SGD) step then reads

$$x^{(n+1)} = x^{(n)} - \alpha^{(n)} \nabla f_{i_n}(x^{(n)})$$

where i_n is picked uniformly at random in $\{1, \ldots, k\}$. As before we have

$$\mathbb{E}_{i_n}\left[\nabla f_{i_n}(x)\right] = \nabla \mathcal{E}(x).$$

Theorem 4.8. *Assume $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}$, $\mathcal{E}(x) = \frac{1}{k}\sum_{i=1}^{k} f_i(x)$ is smooth with L-Lipschitz continuous gradient. Further assume that $\mathcal{X}^* = \arg\min_{x \in \mathbb{R}^d} \mathcal{E}(x) \neq \emptyset$ and that $\mathcal{E}$ fulfills the Polyak–Łojasiewicz inequality with $\mu > 0$ and that it holds*

$$\mathbb{E}_i\left[\|\nabla f_i(x^{(n)})\|^2\right] \leq C^2$$

for all iterates $x^{(n)}$ and for some $C > 0$. We consider the stochastic gradient descent algorithm

$$x^{(n+1)} = x^{(n)} - \alpha^{(n)} \nabla f_{i_n}(x^{(n)}). \tag{4.6}$$

Using the step size

$$\alpha^{(n)} = \frac{2n + 1}{2\mu(n + 1)^2} \to 0 \quad \text{as} \quad n \to \infty$$

gives the convergence rate

$$\mathbb{E}\left[\mathcal{E}(x^{(n)}) - \mathcal{E}(x^*)\right] \leq \frac{LC^2}{4\mu^2} \cdot \frac{1}{n}.$$

Whereas using the step size

$$\alpha^{(n)} \equiv \alpha < \frac{1}{2\mu}$$

gives the convergence rate

$$\mathbb{E}\left[\mathcal{E}(x^{(n)}) - \mathcal{E}(x^*)\right] \leq (1 - 2\mu\alpha)^n \left(\mathcal{E}(x^{(0)}) - \mathcal{E}(x^*)\right) + \frac{LC^2\alpha}{4\mu}.$$

Proof. □

4.2 Hierarchy of properties such that gradient descent converges

In the following we will consider different conditions on a smooth function $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}$ such that gradient descent will converge. We assume that $\mathcal{X}^* = \arg\min_{x \in \mathbb{R}^d} \mathcal{E}(x) \neq \emptyset$ and denote by

$$x_p = \arg\min_{z \in \mathcal{X}^*} \|z - x\|_2^2$$

the projection of x onto the solution set $\mathcal{X}^*$.

- **Strong Convexity (SC)**: There exists some $\mu > 0$ such that for all $x, y \in \mathbb{R}^d$ it holds

$$\mathcal{E}(y) \geq \mathcal{E}(x) + \langle \nabla\mathcal{E}(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2.$$

- **Essential Strong Convexity (ESC)**: There exists some $\mu > 0$ such that for all $x, y \in \mathbb{R}^d$ with $x_p = y_p$ it holds

$$\mathcal{E}(y) \geq \mathcal{E}(x) + \langle \nabla\mathcal{E}(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2.$$

- **Weak Strong Convexity (WSC)**: There exists some $\mu > 0$ such that for all $x \in \mathbb{R}^d$ it holds

$$\mathcal{E}(x_p) \geq \mathcal{E}(x) + \langle \nabla\mathcal{E}(x), x_p - x\rangle + \frac{\mu}{2}\|x_p - x\|^2.$$

- **Restricted Secant Inequality (RSI)**: There exists some $\mu > 0$ such that for all $x \in \mathbb{R}^d$ it holds

$$\langle \nabla\mathcal{E}(x), x - x_p\rangle \geq \mu\|x - x_p\|^2.$$

- **Error Bound (EB)**: There exists some $\mu > 0$ such that for all $x \in \mathbb{R}^d$ it holds

$$\|\nabla\mathcal{E}(x)\| \geq \mu\|x - x_p\|.$$

- **Polyak–Łojasiewicz Inequality (PL)**: There exists some $\mu > 0$ such that for all $x \in \mathbb{R}^d$ it holds

$$\frac{1}{2}\|\nabla\mathcal{E}(x)\|^2 \geq \mu\left(\mathcal{E}(x) - \mathcal{E}(x_p)\right).$$

Theorem 4.9. *For a smooth function $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}$ with L-Lipschitz continuous gradient the following implications are true:*

$$(SC) \implies (ESC) \implies (WSC) \implies (RSI) \implies (EB) \iff (PL) \implies (QG).$$

Proof. □

5 Nonsmooth and nonconvex objective functions

Definition 5.1 (subgradient and subdifferential)**.** *let $\mathcal{E}:\mathbb{R}^d\to\mathbb{R}$ be convex. a vector $p\in\mathbb{R}^d$ is a subgradient of $\mathcal{E}$ at the point $x\in\mathbb{R}^d$ if*

$$\mathcal{E}(y)\geq\mathcal{E}(x)+\langle p,y-x\rangle,\quad\forall y\in\mathbb{R}^d$$

the subdifferential of $\mathcal{E}$ at the point $x\in\mathbb{R}^d$ is given by the set of all subgradients

$$\partial\mathcal{E}(x)=\{p\in\mathbb{R}^d|\mathcal{E}(y)\geq\mathcal{E}(x)+\langle p,y-x\rangle,\forall y\in\mathbb{R}^d\}$$

Theorem 5.2. *let $\mathcal{E}:\mathbb{R}^d\to\mathbb{R}$ be convex. then $\partial\mathcal{E}(x)$ is nonempty, convex and compact for every $x\in\mathbb{R}^d$. furthermore if $\mathcal{E}$ is differentiable at the point $x\in\mathbb{R}^d$ then it holds $\partial\mathcal{E}(x)=\{\nabla\mathcal{E}(x)\}$*

Lemma 5.3. *let $\mathcal{E}:\mathbb{R}^d\to\mathbb{R}$ be convex and $x_1,x_2\in\mathbb{R}^d$ be given. if $p_1\in\partial\mathcal{E}(x_1)$ and $p_2\in\partial\mathcal{E}(x_2)$ then it holds*

$$\langle x_1-x_2,p_1-p_2\rangle\geq 0$$

we now focus on the following generalization of the gradient flow: find a trajectory $t\to x(t)$ such that

$$\begin{cases}\dot{x}(t)\in-\partial\mathcal{E}(x(t))\text{ for a.e. }t\geq 0,\\ x(0)=x_0\end{cases}\tag{DI}$$

(this is called a *differential inclusion*, which is a generalization of ode). We would like to prove existence and uniqueness of solutions for (DI) under the additional assumption that the trajectory is absolutely continuous (AC), i.e., there exists some $\ell\in L^1([0,T])$ such that

$$\|x(t_1)-x(t_2)\|\leq\int_{t_1}^{t_2}\ell(s)\,ds\quad\forall t_1,t_2\in[0,T].\tag{AC}$$

Note that every absolutely continuous function is continuous and it is differentiable almost everywhere (a.e.).
For later use we denote with

$$\partial_0\mathcal{E}(x)=\arg\min_{p\in\partial\mathcal{E}(x)}\|p\|_2$$

the element with minimal norm of $\partial\mathcal{E}(x)$ (since $\partial\mathcal{E}(x)$ is nonempty, compact and convex by Theorem 5.2 and $p\mapsto\|p\|_2^2$ is strictly convex, there exists a unique element in the subdifferential with minimal norm).

Proposition 5.4. *let $\mathcal{E}:\mathbb{R}^d\to\mathbb{R}$ be convex and let x_1 and x_2 be two solutions of DI. then it holds*

$$\|x_1(t)-x_2(t)\|\leq\|x_1(0)-x_2(0)\|\quad\forall t\geq 0$$

in particular if it holds $x_1(0)=x_2(0)$ then $x_1(t)=x_2(t)$ for every $t\geq 0$

Definition 5.5 (Semi-convexity)**.** *The function $\mathcal{E}:\mathbb{R}^d\to\mathbb{R}$ is semi-convex or λ-convex, if there exists a $\lambda\in\mathbb{R}$ such that*

$$x\mapsto\mathcal{E}(x)-\frac{\lambda}{2}\|x\|_2^2\text{ is convex.}$$

Definition 5.6. *Let $\mathcal{E}:\mathbb{R}^d\to\mathbb{R}$ be λ-convex for some $\lambda\in\mathbb{R}$. Then the subdifferential of $\mathcal{E}$ is defined as*

$$\partial\mathcal{E}(x)=\left\{p\in\mathbb{R}^d\,\middle|\,\mathcal{E}(y)\geq\mathcal{E}(x)+\langle p,y-x\rangle+\frac{\lambda}{2}\|y-x\|^2\quad\forall y\in\mathbb{R}^d\right\}.\tag{5.4}$$

If $\mathcal{E}$ is λ-convex with $\lambda\geq 0$ then $\mathcal{E}$ is convex and by Theorem 5.2 the subdifferential is nonempty, compact and convex. For $\lambda<0$ we have

$$\partial\mathcal{E}(x)=\left\{p\in\mathbb{R}^d\,\middle|\,p-\lambda x\in\partial\tilde{\mathcal{E}}(x)\right\},$$

where $\tilde{\mathcal{E}}(x)=\mathcal{E}(x)-\frac{\lambda}{2}\|x\|^2$. Since $\tilde{\mathcal{E}}$ is convex we have that $\partial\tilde{\mathcal{E}}(x)$ is nonempty, compact and convex by Theorem 5.2. This implies that also $\partial\mathcal{E}(x)$ is nonempty, compact and convex and $\partial_0\mathcal{E}(x)$ is well-defined.

Lemma 5.7. *Let $\mathcal{E}:\mathbb{R}^d\to\mathbb{R}$ be λ-convex for some $\lambda\in\mathbb{R}$ and $x_1,x_2\in\mathbb{R}^d$ be given. If $p_1\in\partial\mathcal{E}(x_1)$ and $p_2\in\partial\mathcal{E}(x_2)$ then it holds*

$$\langle x_1-x_2,p_1-p_2\rangle\geq\lambda\|x_1-x_2\|^2.$$

Proposition 5.8. *Let $\mathcal{E}:\mathbb{R}^d\to\mathbb{R}$ be λ-convex for some $\lambda\in\mathbb{R}$ and let x_1 and x_2 be two solutions of (DI). Then it holds*

$$\|x_1(t)-x_2(t)\|\leq\|x_1(0)-x_2(0)\|e^{-\lambda t}\quad\forall t\geq 0.$$

In particular if it holds $x_1(0)=x_2(0)$ then $x_1(t)=x_2(t)$ for every $t\geq 0$.

Proposition 5.9. *Let $\mathcal{E}:\mathbb{R}^d\to\mathbb{R}$ be λ-convex for some $\lambda\in\mathbb{R}$ and let x be a solution of*

$$\begin{cases}\dot{x}(t)\in-\partial\mathcal{E}(x(t)) \text{ for a.e. } t\geq 0,\\ x(0)=x_0,\\ t\mapsto x(t) \text{ is (AC) in } [0,T].\end{cases}$$

Then for any time $t_0\in[0,T]$ such that both $t\mapsto x(t)$ and $t\mapsto\mathcal{E}(x(t))$ are differentiable in $t=t_0$, the subdifferential in $t=t_0$ is contained in the hyperplane orthogonal to $\dot{x}(t_0)$. In particular it holds

$$\dot{x}(t_0)=-\partial_0\mathcal{E}(x(t_0)).$$

Moreover if $t\mapsto\mathcal{E}(x(t))$ is differentiable a.e. then

$$\dot{x}(t)=-\partial_0\mathcal{E}(x(t))$$

for a.e. $t\in[0,T]$.

So far we have proven that if there exists an absolutely continuous solution of (DI) then the solution is unique if the starting point is the same. Now we want to prove existence of such solutions.

Definition 5.10 (minimizing movement scheme, MMS)**.** *For any time step $\tau>0$ we consider the sequence of time points $(t_k^\tau)_{k\in\mathbb{N}}$ with $t_k^\tau=k\tau$ and the sequence of points $(x_k^\tau)_{k\in\mathbb{N}}$ with $x_k^\tau=x(t_k^\tau)$ defined by the de Giorgi minimizing movement scheme*

$$\begin{cases}x_{k+1}^\tau\in\arg\min_{x\in\mathbb{R}^d}\left[\mathcal{E}(x)+\frac{\|x-x_k^\tau\|^2}{2\tau}\right],\\ x_0^\tau=x_0\end{cases}\tag{MMS}$$

We assume that at every time step there exists a x_{k+1}^τ in (MMS). If $\mathcal{E}$ is λ-convex then we have

$$\mathcal{E}(x)+\frac{\|x-x_k^\tau\|^2}{2\tau}=\mathcal{E}(x)-\frac{\lambda}{2}\|x\|^2+\frac{\lambda}{2}\|x\|^2+\frac{\|x-x_k^\tau\|^2}{2\tau}.$$

If $\tau>0$ is small enough such that $\frac{\lambda}{2}+\frac{1}{2\tau}>0$ then the above function is strongly convex and therefore there exists a unique x_{k+1}^τ in (MMS).

Proposition 5.11 (Existence of solutions of (DI))**.** *Let $\mathcal{E}:\mathbb{R}^d\to(-\infty,\infty]$ be given. Further let the two curves x^τ and $\tilde{x}^\tau$ be constructed by using the interpolation of (MMS) as above and suppose $\mathcal{E}(x_0)<+\infty$. Then up to a subsequence of $(\tau_n)_n\to 0$ both $(x^\tau)_\tau$ and $(\tilde{x}^\tau)_\tau$ converge uniformly to the same curve $x\in H^1([0,T])$. Further v^τ converges weakly in $L^2([0,T])$ to a vector-valued function v, such that $\dot{x}=v$. Additionally it holds:*

1. *If $\mathcal{E}$ is λ-convex for some $\lambda\in\mathbb{R}$ then it holds $v(t)\in-\partial\mathcal{E}(x(t))$ for almost every t, i.e., $x(t)$ is a solution of (DI).*
2. *If $\mathcal{E}$ is λ-convex for some $\lambda\in\mathbb{R}$ and additionally $\mathcal{E}\in C^1(\mathbb{R}^d)$ then it holds $v(t)=-\nabla\mathcal{E}(x(t))$ for every t, i.e., $x(t)$ is a gradient descent trajectory.*

So far we have shown existence and uniqueness of a solution of (DI). Now we want to know where such a solution ends up. Let $\mathcal{E}\in C^1(\mathbb{R}^d)$ and x^* be a critical point of $\mathcal{E}$. Let us assume that locally around x^* the Polyak-Łojasiewicz inequality is fulfilled.

we show that for λ-convex functions fulfilling a local Polyak-Łojasiewicz condition the iterations will inevitably get stuck at a local minimizer.

6 Primer on stochastic calculus and stochastic differential equations

Definition 6.1 (σ-algebra)**.** *A σ-algebra $\mathcal{F}$ on a set Ω is a collection of subsets of Ω such that:*

1. *$\Omega\in\mathcal{F}$,*
2. *If $A\in\mathcal{F}$, then $A^c\in\mathcal{F}$,*
3. *If $A_1,A_2,A_3,\dots\in\mathcal{F}$, then $\bigcup_{i=1}^\infty A_i\in\mathcal{F}$.*

The pair $(\Omega,\mathcal{F})$ is called a measurable space.

Definition 6.2 (Filtration)**.** *A filtration $\{\mathcal{F}_t\}_{t\geq 0}$ is an increasing family of σ-algebras such that $\mathcal{F}_s\subseteq\mathcal{F}_t$ for all $0\leq s\leq t$. Each $\mathcal{F}_t$ represents the information available up to time t.*

Definition 6.3 (Random Variable). *A random variable is a measurable function $X : \Omega \to \mathbb{R}$ such that for every Borel set $B \in \mathcal{B}(\mathbb{R})$, the preimage $X^{-1}(B) \in \mathcal{F}$, where $(\Omega, \mathcal{F}, P)$ is a probability space.*

Definition 6.4 (Gaussian Random Variable). *A Gaussian random variable is a random variable X such that for some $\mu \in \mathbb{R}$ and $\sigma^2 \geq 0$, the probability density function is given by*

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

Definition 6.5 (Stochastic Process). *A stochastic process $\{X_t\}_{t\geq 0}$ is a collection of random variables indexed by time $t \in T$ (often $T = [0, \infty)$), such that for each t, X_t is defined on a common probability space $(\Omega, \mathcal{F}, P)$.*

Definition 6.6 (Brownian Motion). *A Brownian motion (or Wiener process) A standard one-dimensional Brownian motion $\{W_t\}_{t\geq 0}$ is an adapted real-valued and continuous stochastic process satisfying:*

1. *$W_0 = 0$ (almost surely),*
2. *W_t has independent increments, ($W_t - W_s$ is an independent random variable w.r.t. W_r for all $r \leq s$)*
3. *$W_t - W_s \sim N(0, t-s)$ for $0 \leq s < t$, ($W_t \sim \mathcal{N}(0, t)$)*

Proof. Existence of Brownian motions (by construction) □

Proposition 6.7. *properties of Brownian motion:*

1. $\mathbb{E}[W(t)] = 0$
2. $Var[W(t)] = t = \mathbb{E}[W^2(t)]$
3. $Cov[W(s), W(t)] = s \wedge t$

Proof. assume $0 < s < t$:

$$Cov[W(s), W(t)] = \mathbb{E}[W(s)W(t)] - \mathbb{E}[W(s)]\mathbb{E}[W(t)] = \mathbb{E}[W(s)W(t)] - 0 = \mathbb{E}\{W(s)[W(t) - W(s) + W(s)]\}$$
$$= \mathbb{E}\{W(s)[W(t) - W(s)]\} + \mathbb{E}[W^2(s)]$$

by independence of increments, $\mathbb{E}\{W(s)[W(t) - W(s)]\} = 0$, we have

$$Cov[W(s) - W(t)] = \mathbb{E}[W^2(s)] = s$$

□

Example 6.8. *assume $0 < s < t$, calculate mean and variance of $W(s) - W(t)$:*

solution. $W(s) + W(t) = 2W(s) + [W(t) - W(s)]$,

$$Var[W(s) - W(t)] = Var[2W(s) + [W(t) - W(s)] = 4Var[W(s)] + Var[W(t) - W(s)]$$
$$= 4Var[W(s)] + (t - s) = 2s + (t - s) = 3s + t$$

we use the independence of increments ($\mathbb{E}\{W(s)[W(s) - W(t)]\} = 0$). □

consider the so-called instantaneous increment: $dW(t) = \lim_{\Delta t \to 0} W(t + \Delta t) - W(t)$, taking derivative we have

$$\frac{dW(t)}{dt} = \lim_{\Delta t \to 0} \frac{W(t + \Delta t) - W(t)}{\Delta t}$$

According to the properties of Brownian motion:

$$\mathbb{E}\left[\frac{W(t + \Delta t) - W(t)}{\Delta t}\right] = \frac{1}{\Delta t} \cdot \mathbb{E}[W(t + \Delta t) - W(t)] = 0$$

$$\text{Var}\left[\frac{W(t + \Delta t) - W(t)}{\Delta t}\right] = \frac{1}{(\Delta t)^2} \cdot \text{Var}[W(t + \Delta t) - W(t)] = \frac{1}{\Delta t}$$

When $\Delta t \to 0$, $\text{Var}\left[\frac{W(t+\Delta t)-W(t)}{\Delta t}\right] \to \infty$, which means that the value of the derivative can be arbitrarily large. Thus, it can be concluded that the derivative of $W(t)$ does not exist.

Definition 6.9 (Square-Integrable Process). *A stochastic process $\{X_t\}_{t\geq 0}$ is square-integrable if*

$$\|X\|_{\mathbb{Q}}^2 \triangleq \int_0^\infty \mathbb{E}[X_t^2]dt < +\infty \quad \textit{for all } t \geq 0.$$

$\mathbb{Q}$ is the space of square-integrable stochastic processes.

Definition 6.10 (Simple Process). *A simple process $Y_t(\omega)$ is a stochastic process $\in \mathbb{Q}$ of the form*

$$Y_t(\omega) = \sum_{i=0}^{+\infty} \xi_i(\omega)\mathbf{1}_{[t_i,t_i+1)}(t),$$

where $0 = t_0 < t_1 < \cdots < t_n \to \infty$ is a partition of $[0,T]$, and ξ_i are independent random variables.

For simple processes $Y_t(\omega)$ we define

$$\int_0^{+\infty} Y_t(\omega)\, dW_t(\omega) := \sum_{i=0}^{+\infty} \eta_i \left(W_{t_{i+1}}(\omega) - W_{t_i}(\omega)\right). \tag{6.1}$$

Since this sum is an infinite series we need to justify the convergence of this series and we do it in $L^2(\Omega, P)$ which is the space of square-integrable functions X such that

$$\int_\Omega |X(\omega)|^2\, dP(\omega) \left(= \int_{\mathbb{R}} |\xi|^2 \rho_X(\xi)\, d\xi\right) < +\infty.$$

Proposition 6.11. *The sum in* (6.1) *converges in $L^2(\Omega, P)$. Moreover it holds*

$$\mathbb{E}\left[\left(\int_0^{+\infty} Y_t\, dW_t\right)^2\right] = \int_0^{+\infty} \mathbb{E}[Y_t^2]\, dt = \|Y\|_{\mathbb{Q}}^2.$$

This identity is called Itô isometry for simple processes.

Proof. □

Proposition 6.12. *Simple processes are dense in $\mathbb{Q}$: for every $Y \in \mathbb{Q}$ there exist $Y_n, n \in \mathbb{N}$ simple processes s.t.*

$$\lim_{n\to\infty} \|Y_n - Y\|_{\mathbb{Q}} = 0$$

Proof. □

Definition 6.13 (Itô Integral). *The Itô integral of a stochastic process $Y(t,\omega) \in \mathbb{Q}$ with simple processes $Y_n(t,\omega) \to Y(t,\omega)$ with respect to a Brownian motion W_t is defined as*

$$\int_0^\infty Y(t,\omega)dW_t \triangleq \lim_{n\to\infty} \int_0^\infty Y_n(t,\omega)dW_t$$

For more general integrands, the Itô integral is defined as the limit of the integrals for simple processes.

Definition 6.14 (Itô Isometry). *The Itô isometry states that for any square-integrable process $Y_t \in \mathbb{Q}$,*

$$\mathbb{E}\left[\left(\int_0^\infty Y_t\, dW_t\right)^2\right] = \int_0^\infty \mathbb{E}[Y_t^2]\, dt.$$

Definition 6.15 (Quadratic Variation). *let $f : \mathbb{R}_+ \to \mathbb{R}$ be given and consider a partition $\Pi = \{t_i, i \in \mathbb{N}\}$ with $t_0 = 0, t_i < t_{i+1}$, $t_i \to +\infty$ and denote $|\Pi| = \max_{i\in\mathbb{N}} |t_{i+1} - t_i|$ the size of the partition Π. the quadratic variation is defined as*

$$\langle f\rangle_t = \lim_{|\Pi|\to 0} Q(f,\Pi)_t$$

where $Q(f,\Pi)_t = \sum_{t_i\in\Pi} |f(t_{i+1}\wedge t) - f(t_i \wedge t)|^2$. ($a \wedge b \triangleq \min(a,b)$)

Example 6.16. *for smooth function f, $\langle f\rangle_t = 0, \quad \forall t \in \mathbb{R}_+$*

Proof. □

Theorem 6.17. *for every fixed $\omega \in \Omega$, the identity*

$$\langle W(\cdot,\omega)\rangle_t = t$$

holds in $L^2(\Omega,\mathbb{P})$

Proof. □

Theorem 6.18 (Itô formula (integral form))**.** *For $h \in \mathbb{Q}$ and $\{g_t, t \geq 0\}$ adapted to the filtration such that*

$$\int_0^{+\infty} |g(t)|\, dt < +\infty$$

a.s., i.e.,

$$\mathbb{P}\left(\left\{\omega \,\middle|\, \int_0^{+\infty} |g(t,\omega)|\, dt < +\infty\right\}\right) = 1,$$

then for all $t \geq 0$

$$X(t,\omega) = \int_0^t g(s,\omega)\, ds + \int_0^t h(s,\omega)\, dB_s(\omega)$$

is an adapted stochastic process.

If we consider $\varphi \in C_b^{2,1}(\mathbb{R},\mathbb{R})$, twice continuously differentiable with Lipschitz-continuous and bounded Hessian and $Y_t = \varphi(X_t)$, then the Itô formula holds:

$$Y_t = Y(t,\omega) = Y_0 + \int_0^t \frac{\partial\varphi}{\partial X}(X_s)g(s)\, ds + \int_0^t \frac{\partial\varphi}{\partial X}(X_s)h(s)\, dB_s + \frac{1}{2}\int_0^t \frac{\partial^2\varphi}{\partial X^2}(X_s)h(s)^2\, ds. \tag{6.4}$$

Proof. □

Notation: In the following we will write

$$dX(t) = g(t)dt + h(t)dB_t$$

and mean

$$X(t) = X(0) + \int_0^{+\infty} g(s)\, ds + \int_0^t h(s)\, dB_s$$

in $L^2(\Omega,\mathbb{P})$.

We consider

$$dX(t) = V(t)dt + M(t)dB_t \tag{6.5}$$

where

- $t \mapsto V(t) \in \mathbb{R}^n$ is sufficiently integrable,
- $t \mapsto M(t) \in \mathbb{R}^{n\times m}$ is sufficiently integrable and
- $t \mapsto B_t \in \mathbb{R}^m$ is a coordinate-wise Brownian motion

so that $t \mapsto X(t) \in \mathbb{R}^n$ is a vector-valued stochastic process.

Theorem 6.19 (Itô formula (vector form))**.** *Fix a vector-valued stochastic process as in* (6.5) *and let $\varphi : \mathbb{R}^n \times \mathbb{R}_+ \to \mathbb{R}^p$ be a function which is twice continuously differentiable in space with Lipschitz-continuous Hessian and continuously differentiable in time. Then $Y_t = \varphi(X_t, t)$ is the p-dimensional stochastic process with k-th component given by*

$$dY_k(t) = \frac{\partial\varphi_k}{\partial t}dt + \sum_{i=1}^n \frac{\partial\varphi_k}{\partial X_i}dX_i + \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \frac{\partial^2\varphi_k}{\partial X_i\partial X_j}\, dX_i \cdot dX_j. \tag{6.6}$$

From (6.5) *we have*

$$dX_i = V_i dt + \sum_{j=1}^m M_{ij}dB_j.$$

To give a meaning to

$$dX_i \cdot dX_j,$$

we compute:

$$dX_i \cdot dX_j = \left(V_i dt + \sum_{k=1}^{m} M_{ik} dB_k\right) \cdot \left(V_j dt + \sum_{l=1}^{m} M_{jl} dB_l\right).$$

Expanding this, we have:

$$dX_i \cdot dX_j = V_i V_j dt^2 + \sum_{l=1}^{m} V_i M_{jl} dt dB_l + \sum_{k=1}^{m} V_j M_{ik} dt dB_k + \sum_{k=1}^{m}\sum_{l=1}^{m} M_{ik} M_{jl} dB_k dB_l.$$

We define

$$dB_k \cdot dB_l = \begin{cases} dt & \text{if } k = l, \\ 0 & \text{if } k \neq l. \end{cases}$$

Then we have

$$dX_i dX_j = \sum_k \sum_l M_{ik} M_{jl} dB_k dB_l = \sum_k M_{ik} M_{jk} \delta_{kl} dt = \left[\sum_k M_{ik} M_{jk}\right] dt.$$

And

$$\frac{1}{2} \sum_{i,j} \frac{\partial \varphi_k}{\partial X_i \partial X_j} dX_i dX_j = \frac{1}{2} \operatorname{tr}(M^\top \nabla^2 \varphi_k M) dt.$$

Example 6.20 (2-dimensional ito formula)**.** *let first and second order derivatives of $f(t,x,y)$ exist and be continuous. $X(t)$ and $Y(t)$ are ito processes. we have*

$$df = f_t dt + f_x dX(t) + f_y dY(t) + f_{xy}[dX(t)dY(t)] + \frac{1}{2} f_{xx}[dX(t)]^2 + \frac{1}{2} f_{yy}[dY(t)]^2$$

Example 6.21 (ito product rule)**.** *for $X(t)$ and $Y(t)$ be ito process, $d[X(t)Y(t)]$ is computed by ito product rule:*

$$d[X(t)Y(t)] = X(t)dY(t) + Y(t)dX(t) + dX(t)dY(t)$$

Definition 6.22 (Stochastic differential equation)**.** *we define stochastic processes as implicit solutions of*

$$dX_t = g(t, X_t)dt + h(t, X_t)dB_t$$

under the following conditions we can verify the existence and uniqueness of such solutions:

- *local lipschitz continuity: there exists $R > 0$ and a constant $C_R > 0$ such that for every $x, y \in \mathbb{B}(0, R)$ it holds:*

$$|g(t,x) - g(t,y)| \leq C_R |x - y|, \quad |h(t,x) - h(t,y)| \leq C_R |x - y|$$

- *sublinear growth: there exists a constant $C > 0$ such that for every $x \in \mathbb{R}^d$ it holds*

$$|g(t,x)| \leq C(1 + |x|), \quad |h(t,x)| \leq C(1 + |x|)$$

Proof.

- uniqueness
- existence

□

Example 6.23. *Given the $\mathcal{F}(t)$-measurable stochastic process $X(t)$, its expression is as follows:*

$$X(t) = \exp\left[\theta W(t) - \frac{1}{2}\theta^2 t\right],$$

where $W(t)$ is the standard Brownian motion, and θ is a constant. find the SDE of this stochastic process.

solution. let $Y(t) = \theta W(t) - \frac{1}{2}\theta^2 t, \quad X(t) = \varphi(Y(t)) = \exp[Y(t)]$, by Ito's formula:

$$dX(t) = \frac{\partial \varphi(Y_t)}{\partial t}dt + \frac{\partial \varphi(Y_t)}{\partial Y_t}dY_t + \frac{1}{2}\frac{\partial^2 \phi(Y_t)}{\partial Y_t^2}dY_t^2$$

$$= X(t)[\theta dW_t - \frac{1}{2}\theta^2 dt] + \frac{1}{2}X(t)[\theta dW_t - \frac{1}{2}\theta^2 dt]^2 = X(t)[\theta dW_t - \frac{1}{2}\theta^2 dt] + \frac{1}{2}X(t)\theta^2 dt = X(t)\theta dW_t$$

□

Example 6.24 (Geometric Brownian Motion (GBM))**.** *solve the SDE:*

$$dS(t) = \mu S(t)dt + \sigma S(t)dW(t), \quad S(0) = S_0$$

where $\mu(t)$ and $\sigma(t)$ are continuous bounded functions.

solution.

$$\frac{dS(t)}{S(t)} = \mu dt + \sigma dW(t) \Rightarrow d\ln S(t) = \mu dt + \sigma dW(t)$$

by ito formula:

$$dY_k(t) = \frac{\partial \varphi_k}{\partial t}dt + \sum_{i=1}^{n}\frac{\partial \varphi_k}{\partial X_i}dX_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial^2 \varphi_k}{\partial X_i \partial X_j}dX_i \cdot dX_j. \tag{6.6}$$

let $n = 1$ and $k = 1$, this can be written as

$$dY(t) = \frac{\partial \varphi(S_t)}{\partial t}dt + \frac{\partial \varphi(S_t)}{\partial S_t}dS_t + \frac{1}{2}\frac{\partial^2 \phi(S_t)}{\partial S_t^2}dS_t^2$$

let $\phi(S_t) = \ln S_t = \ln S(t)$, $g(t) = \mu$ and $h(t) = \sigma$, we have

$$d\varphi(S_t) = \frac{\partial \varphi(S_t)}{\partial t}dt + \frac{\partial \varphi(S_t)}{\partial S_t}dS_t + \frac{1}{2}\frac{\partial^2 \varphi(S_t)}{\partial S_t^2}(dS_t)^2 = 0 + \frac{1}{S_t}[\mu S_t dt + \sigma S_t dW_t] + \frac{1}{2}(-\frac{1}{S_t^2})[\mu S_t dt + \sigma S_t dW_t]^2$$

we have $[\mu S_t dt + \sigma S_t dW_t]^2 = (\sigma S_t)^2 dW_t^2 = (\sigma S_t)^2 dt$ since $(dt)^2 = 0 = dt \cdot dW_t$ and $dW_t^2 = dt$

$$d\varphi(S_t) = \mu dt + \sigma dW_t - \frac{1}{2S_t^2}(\sigma S_t)^2 dt = (\mu - \frac{1}{2}\sigma^2)dt + \sigma dW_t$$

$$\ln S(t) - \ln S(0) = \int_0^t (\mu - \frac{1}{2}\sigma^2)dt + \int_0^t \sigma dW_t = (\mu - \frac{1}{2}\sigma^2)t + \sigma[W(t) - W(0)]$$

$$S(t) = S_0 \exp\left[(\mu - \frac{1}{2}\sigma^2)t + \sigma W(t)\right]$$

□

7 Simulated Annealing

the roadmap of this chapter is:

1. the iteration scheme of SGD is similar to the Euler-Maruyama scheme, which converges to the solution of the SDE Langevin equation:
$$dX_t = -\nabla \mathcal{E}(X_t)dt + \sqrt{2\sigma}dB_t$$
2. we study the behavior of the trajectory of Langevin equation's density $\rho_t \in \mathcal{P}(\mathbb{R}^d)$
3. we show that ρ_t satisfies the Fokker-Planck equation (FPE):
$$\frac{\partial}{\partial t}\rho_t(x) = \operatorname{div}(\nabla \mathcal{E}(x)\rho_t(x)) + \sigma \Delta \rho_t(x)$$

4. we show that FPE can be interpreted as the gradient descent of a suitable function J on $\mathcal{P}_{AC}(\mathbb{R}^d)$:

$$J : \mathcal{P}_{AC}(\mathbb{R}^d) \to \mathbb{R}, \quad J(\rho) \triangleq J_\sigma(\rho) = \underbrace{\int_{\mathbb{R}^d} \mathcal{E}(x)\rho(x)dx}_{:=G(\rho)} + \sigma \underbrace{\int_{\mathbb{R}^d} \log(\rho(x))\rho(x)dx}_{:=H(\rho)}$$

J is a functional on $\mathcal{P}_2(\mathbb{R}^d)$ with a metric called Wasserstein distance $W_2(\rho, \rho')$. here gradient descent/gradient flow is defined as:

$$\frac{\partial \rho(t)}{\partial t} = -\nabla J(\rho(t))$$

with Wasserstein distance, gradient descent/flow can be written as:

$$\frac{\partial \rho}{\partial t} = \nabla \cdot \left(\rho \nabla_x \frac{\partial J(\rho)}{\partial \rho}(x) \right)$$

where $\frac{\partial J(\rho)}{\partial \rho}(x)$ is called Variational Derivative of $J(\rho)$

5. in gradient descent one stops at critical points or steady states. we next study the steady state of $\bar{\rho}$ of FPE. $\bar{\rho}$ is characterized as the zero solution of the RHS of FPE:

$$0 = \operatorname{div}(\nabla \mathcal{E} \bar{\rho}) + \sigma \Delta \bar{\rho}$$

we show that $\bar{\rho} = \bar{\rho}_\sigma = \arg\min_{\rho \in \mathcal{P}_2} J_\sigma(\rho)$, here $\bar{\rho}$ is called the Gibbs density associated to $\frac{\mathcal{E}}{\sigma}$. if $\rho_t \to \bar{\rho}_\sigma$ for $t \to +\infty$ then it means that the probability that X_t for t large is getting close to x^* is very high, especially for $\sigma \approx 0$. note that if $\sigma = 0$, it is GD. and we know from former chapter that X_t will not converge to x^* unless $\mathcal{E}(x)$ fulfills conditions like the PL condition. Now we have $\sigma > 0$ and we wonder what condition we can have for J_σ such that the solution of FPE, ρ_t converges to the global minimizer $\bar{\rho}_\sigma$.

recall PL Condition: The square of the gradient and the difference in objective function values have a linear relationship. As a result, through gradient descent, the objective function value can decrease rapidly. We introduce Log-Sobolev inequality (LSI): Fisher information (analogous to the square of the gradient) and relative entropy have a linear relationship. Consequently, via the gradient flow of the distribution, entropy can decrease rapidly.

6. LSI says that if $\mathcal{E}$ is strongly convex, $\nabla^2 \mathcal{E}(x) \succeq \lambda I$ for some $\lambda > 0$, then for $\nu = \frac{e^{-\mathcal{E}(x)}}{\int_{\mathbb{R}^d} e^{-\mathcal{E}(x)} dx} \in \mathcal{P}_{AC}(\mathbb{R}^d)$:

$$\int_{\mathbb{R}^d} \log(\frac{\mu}{\nu})\mu dx =: \underbrace{H(\mu|\nu)}_{\substack{\text{relative} \\ \text{entropy}}} \leq \frac{1}{2\lambda} \underbrace{I(\mu|v)}_{\substack{\text{Fisher} \\ \text{information}}} := \frac{1}{2}\lambda \int_{R^d} |\nabla \log(\frac{\mu}{\nu})|^2 \mu dx, \quad \forall \mu \in \mathcal{P}_{AC}(\mathbb{R}^d)$$

it has been proved that if ν is a density satisfying LSI(λ) and $\psi \in L^\infty(\mathbb{R}^d)$, then

$$\tilde{\nu} = e^{-\psi(x)}\nu(x) \in \mathcal{P}_{AC}(\mathbb{R}^d)$$

fulfills the LSI($\tilde{\lambda}$) where $\tilde{\lambda} = \lambda e^{\inf \psi - \sup \psi}$. notice that $\psi(x)$ can be nonconvex! in particular if $\nu(x) = e^{-\mathcal{E}(x)}$ then $\tilde{\nu}(x) = e^{-(\mathcal{E}(x)+\psi(x))}$. similarly we consider $\nu = e^{-\frac{\mathcal{E}+\psi}{\sigma}}$, which fulfills LSI($\lambda_\sigma$) with $\lambda_\sigma = \frac{\lambda}{\sigma} e^{\frac{1}{\sigma}(\inf \psi - \sup \psi)}$. consider the Gibbs density:

$$\bar{\rho}_\sigma = \frac{1}{z_\sigma} e^{-\frac{\tilde{\mathcal{E}}(x)}{\sigma}} \text{ with } z_\sigma = \int_{\mathbb{R}^d} e^{-\frac{\tilde{\mathcal{E}}(x)}{\sigma}} dx$$

then $\bar{\rho}_\sigma$ also fulfills the LSI(λ_σ) where $\lambda_\sigma = \frac{\lambda}{\sigma} e^{\frac{1}{\sigma}(\inf \psi - \sup \psi)}$. This is the new condition for the convergence of FPE.

7. assume $t \to \rho_t$ is the solution of FPE, we show that $\frac{d}{dt} H(\rho_t|\bar{\rho}_\sigma) = -\sigma I(\rho_t|\bar{\rho}_\sigma)$, therefore by LSI($\lambda_\sigma$) we have ρ_t fulfills

$$\frac{d}{dt} H(\rho_t|\bar{\rho}_\sigma) = -\sigma I(\rho_t|\bar{\rho}_\sigma) \leq -2\sigma\lambda_\sigma H(\rho_t|\bar{\rho}_\sigma)$$

by using Gronwall's inequality we conclude that $H(\rho_t|\bar{\rho}_\sigma) \leq H(\rho_0|\bar{\rho}_\sigma)e^{-2\sigma\lambda_\sigma t} \to 0$ as $t \to +\infty$. we use the facts that

- $H(\mu|\nu) = 0 \Leftrightarrow \mu = \nu$
- Talagrand inequality: if ν fulfills LSI(λ) then it holds: $W_2^2(\mu, \nu) \leq \frac{2}{\lambda} H(\mu, \nu)$

so we get $W_2^2(\rho_t, \bar{\rho}_\sigma) \leq \frac{2}{\lambda_\sigma} H(\rho_0|\bar{\rho}_\sigma)e^{-2\sigma\lambda_\sigma t} \to 0$ as $t \to +\infty$

8. next we study the behavior of $\bar{\rho}_\sigma(x) = \frac{1}{z_\sigma} e^{-\frac{\mathcal{E}(x)}{\sigma}}, \quad z_\sigma = \int_{\mathbb{R}^d} e^{-\frac{\mathcal{E}(x)}{\sigma}} dx$. the conclusion is that the $\bar{\rho}_\sigma$-measure of the set of points far from the minimizer of $\mathcal{E}$ is vanishing for $\sigma \to 0$.

9. we have shown that for $\sigma > 0$ fixed, $\bar{\rho}_\sigma$ is concentrated around sets of global minimizers. The stochastics of this system is controlled by σ and we set $\sigma_t \downarrow 0$ as $t \to +\infty$, this is the idea why we call the algorithm simulated annealing (SA).

10. we conclude this chapter by showing the convergence of SA:

    Theorem 7.1. *assume $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}_+$ and that there exist constants $C_1, C_2 > 0$ such that*

    $$\mathcal{E}(x) \geq C_1 \|X\|^2 - C_2 \quad \forall x \in \mathbb{R}^d$$

    further assume that for any $\sigma > 0$, $\bar{\rho}(x) = \frac{1}{z_\sigma} e^{-\frac{\mathcal{E}(x)}{\sigma}}$ fulfills the lo-Sobolev inequality with constant $\lambda_\sigma \geq C_0 e^{-\frac{\alpha^*}{\sigma}}$ for some constants $\alpha^* > 0$ and $C_0 > 0$. assume that $t \to \sigma_t$ is smooth, decreases and for t large it holds $\sigma_t = \frac{\alpha}{\log(t)}$ for some $\alpha > \alpha^*$. let be $\rho_0 \in \mathcal{P}_{AC}(\mathbb{R}^d)$ with $M_2(\rho_0) = \int_{\mathbb{R}^d} |x|^2 \rho_0(x) dx < +\infty$ such that $J_\sigma(\rho_0) < +\infty$. then for every $\epsilon > 0$ there exist $C, \tilde{C} > 0$ such that

    $$J_{\sigma_t}(\rho_t) - J_{\sigma_t}(\bar{\rho}_{\sigma_t}) \leq C t^{-(1-\frac{\alpha^*}{\alpha}-\epsilon)}$$

    and $G(\rho_t) - \inf_\rho G(\rho) \leq \tilde{C} \frac{\log(\log(t))}{\log(t)} \to 0, t \to +\infty$

7.1 Log-Sobolev inequality as Polyak-Lojasiewicz condition for J_σ

7.2 Concentration of the Gibbs measure

7.3 Convergence of Simulated Annealing

8 Laplace principle

9 Particle swarm optimization and consensus-based optimization