Lecture: Optimization in Machine Learning

Simon Weissmann

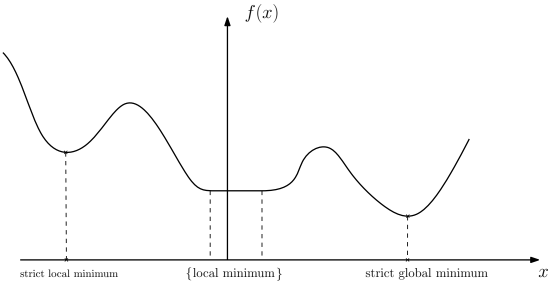
Assistant Professor of Applied Stochastics,
office: B6, 26 – Room B 3.05
email: simon.weissmann@uni-mannheim.de

Thursday, 12:00 - 13:30
December 5, 2024

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Ch2: Unconstrained Optimization methods

Unconstrained Optimization methods



Optimality conditions

Necessary optimality conditions:

Let $f : \mathbb{R}^d \to \mathbb{R}$, $S \subset \mathbb{R}^d$ open, and let $x_* \in S$ be a local minimum of f

- If f is continuously differentiable over S, then $\nabla f(x_*) = 0$.
- If f is twice continuously differentiable over S, then $\nabla^2 f(x_*)$ is positive semi-definite.

Optimality conditions

Sufficient optimality conditions:

Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable over open subset $S \subset \mathbb{R}^d$, and $x_* \in S$ with

- $\nabla f(x_*) = 0$.
- $\nabla^2 f(x_*)$ positive definite.

Then x_* is a strict local minimum of f and there exist $\gamma > 0$, $\varepsilon > 0$ such that

$$f(x) \geq f(x_*) + \frac{\gamma}{2}\|x - x_*\|^2$$

for all $x \in \mathcal{B}_\varepsilon(x_*)$.

Optimality conditions

Optimality condition for convex functions

Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and convex.

1. local minimum of $f \Rightarrow$ global minimum of f.
2. f strictly convex $\Rightarrow$ there exists at most one global minimum of f.
3. $\nabla f(x_*) = 0$ sufficient and necessary condition for global minimum of f.

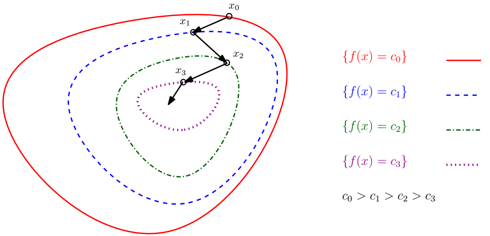
Descent methods

Descent direction

$d \in \mathbb{R}^d$ *descent direction* of f in $x \in \mathbb{R}^d$: $\exists \bar{\alpha} > 0$ such that $f(x + \alpha d) < f(x)$ for all $\alpha \in (0, \bar{\alpha}]$.

Descent condition

$\nabla f(x)^\top d < 0 \quad \Rightarrow \quad d \in \mathbb{R}^d$ descent direction of f in x.



Descent methods

Gradient based methods

$$x_{k+1} = x_k - \alpha_k D_k \nabla f(x_k)$$

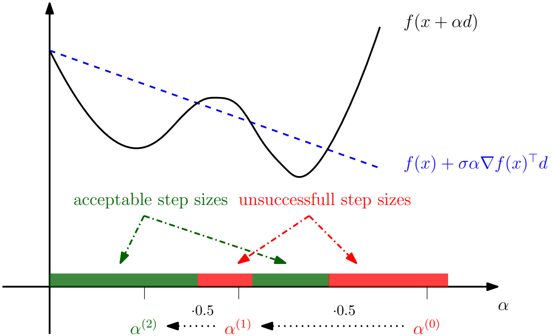
examples:

- *Gradient descent:* $D_k = \mathrm{Id}$
- *Newton method:* $D_k = (\nabla^2 f(x_k))^{-1}$
- *Quasi-Newton method:* $D_k \approx (\nabla^2 f(x_k))^{-1}$

Descent methods

step size selection:

- *Constant step size:* $\alpha_k = s > 0$ for all $k \in \mathbb{N}$
- *Diminishing step size:* $\lim_{k\to\infty} \alpha_k = 0$
- *Armijo rule:*



Convergence of gradient descent

Theorem

- $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable,
- $(x_k)_{k\in\mathbb{N}}$ be generated by

$$x_{k+1} = x_k + \alpha_k d_k, \quad d_k = -\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|},$$

- $\alpha_k > 0$ is chosen by the Armijo step size rule,

Then it holds true that every accumulation point $\bar{x} \in \mathbb{R}^d$ of the sequence $(x_k)_{k\in\mathbb{N}}$ is a stationary point of f, i.e. $\nabla f(\bar{x})$.

Convergence of gradient descent

Definition

$f : \mathbb{R}^d \to \mathbb{R}$ L-smooth, $L > 0 \quad :\Leftrightarrow \quad f$ differentiable & L-Lipschitz gradients, i.e.

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad x, y \in \mathbb{R}^d.$$

Descent Lemma

$f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth: $f(x + y) \leq f(x) + y^\top \nabla f(x) + \frac{L}{2}\|y\|^2$

If $\alpha \leq \frac{2}{L}$:

$$f(x - \alpha \nabla f(x)) \leq f(x) - \alpha\|\nabla f(x)\|^2 + \alpha^2 \frac{L}{2}\|\nabla f(x)\|^2 \leq f(x)$$

Convergence of gradient descent

Theorem (convergence GD with constant step size)

- $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth
- $(x_k)_{k\in\mathbb{N}}$ generated by

$$x_{k+1} = x_k - \bar{\alpha}\nabla f(x_K),$$

  with $\bar{\alpha} \in [\varepsilon, \frac{2-\varepsilon}{L}]$, $\varepsilon \in (0, \frac{2}{L+1})$.

Then every accumulation point $\bar{x} \in \mathbb{R}^d$ of $(x_k)_{k\in\mathbb{N}}$ is a stationary point of f, i.e. $\nabla f(\bar{x}) = 0$.

Convergence of gradient descent

Theorem (convergence GD with diminishing step size)

- $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth
- $(x_k)_{k\in\mathbb{N}}$ generated by

$$x_{k+1} = x_k - \bar{\alpha}\nabla f(x_K),$$

where $\alpha_k > 0$ with

$$\lim_{k\to\infty} \alpha_k = 0 \quad \text{and} \quad \sum_{k=0}^{\infty} \alpha_k = \infty.$$

Then for $(f(x_k))_{k\in\mathbb{N}}$ it holds true that either

$$\lim_{k\to\infty} f(x_k) = -\infty \quad \text{or} \quad \lim_{k\to\infty} \nabla f(x_k) = 0.$$

Moreover, every accumulation point $\bar{x} \in \mathbb{R}^d$ of $(x_k)_{k\in\mathbb{N}}$ is a stationary point of f, i.e. $\nabla f(\bar{x}) = 0$.

Convergence of gradient descent

Theorem (GD convex and smooth)

- $f : \mathbb{R}^d \to \mathbb{R}$ convex and L-smooth with $\inf_x \; f(x) > -\infty$,
- $(x_k)_{k\in\mathbb{N}}$ generated by

$$x_{k+1} = x_k - \bar{\alpha}\nabla f(x_k),$$

  with $\bar{\alpha} \leq \frac{1}{L}$.

Then the sequence $(x_k)_{k\in\mathbb{N}}$ converges in the sense that

$$e(x_k) := f(x_k) - f_* \leq \frac{c}{k}, \quad k \in \mathbb{N}$$

for some constant $c > 0$ and $f_* = \min_{x\in\mathbb{R}^d} f(x)$.

Convergence of gradient descent

$f : \mathbb{R}^d \to \mathbb{R}$ with

- L-smooth $\implies$ $f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|x - y\|^2$
- μ-strongly convex $\implies$ $f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|x - y\|^2$

L-smooth $+$ μ-strongly convex: $\frac{\mu}{2}\|x - y\|^2 \leq f(y) - f(x) - \nabla f(x)^\top (y - x) \leq \frac{L}{2}\|x - y\|^2$

In particular: $\frac{\mu}{2}\|x - x_*\|^2 \leq f(x) - f(x_*) \leq \frac{L}{2}\|x - x_*\|^2$

Convergence of gradient descent

Theorem (GD strong convex and smooth)

- $f : \mathbb{R}^d \to \mathbb{R}$ μ-convex and L-smooth,
- $x_* \in \mathbb{R}^d$ unique global minimum of f, $f(x_*) = \min_{x \in \mathbb{R}^d} f(x)$,
- $(x_k)_{k \in \mathbb{N}}$ generated by
$$x_{k+1} = x_k - \bar{\alpha} \nabla f(x_k),$$
with $\bar{\alpha} = \frac{2}{\mu + L}$.

Then the sequence $(x_k)_{k \in \mathbb{N}}$ converges linearly in the sense that

$$e(x_k) := \|x_k - x_*\| \leq \left(\frac{\kappa - 1}{\kappa + 1} \right)^k \|x_0 - x_*\|, \quad k \in \mathbb{N}$$

where $\kappa = \frac{L}{\mu}$.

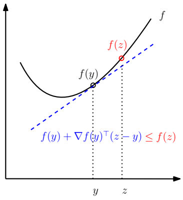
Sub-gradient descent method

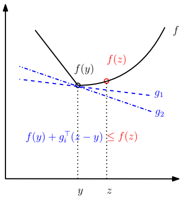
Definition

$g_x \in \mathbb{R}^d$ sub-gradient of $f : \mathbb{R}^d \to \mathbb{R}$ in $x \in \mathbb{R}^d$ if

$$f(y) \geq f(x) + g_x^\top (y - x).$$

for all $y \in \mathbb{R}^d$. Sub-differential: Set of all sub-gradients of f in x of f denoted by $\partial f(x)$.





Sub-gradient descent method

Algorithm: Sub-gradient descent method

- find a sub-gradient $g_{x_k} \in \partial f(x_k)$
- set $x_{k+1} = x_k - \alpha_k g_{x_k}$

Theorem (Sub-gradient descent convergence)

- $f : \mathbb{R}^d \to \mathbb{R}$ be convex and M-Lipschitz continuous,
- $\alpha_k > 0$,
- assume existence of a global minimum $x_* \in \mathbb{R}^d$ of f

Then for $\bar{x}_N := \sum_{k=0}^N w_k x_k$, $w_k = \frac{\alpha_k}{\sum_{s=0}^N \alpha_s}$, $k = 1, \dots, N$ it holds true that

$$e(x_k) = f(\bar{x}_N) - f(x_*) \leq \frac{\|x_0 - x_*\| + M^2 \sum_{k=0}^N \alpha_k^2}{2 \sum_{k=0}^N \alpha_k} .$$

Ch3: Accelerated gradient descent method

Accelerated gradient descent method

Gradient descent struggles with quadratic cost functions of high condition number:

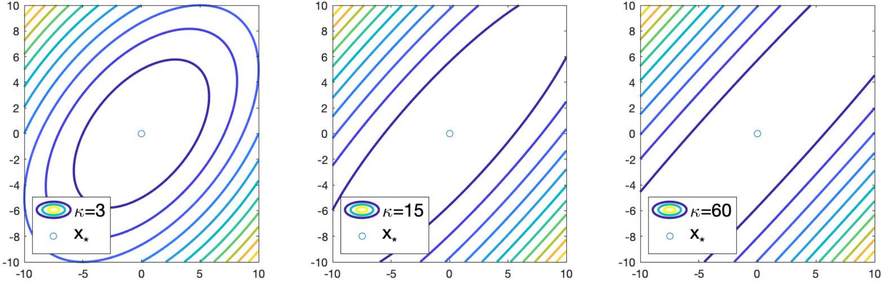


Figure: Contour lines of a quadratic function for increasing condition number κ.

Accelerated gradient descent method

Polyak's heavy ball method (HBM):

$$x_{k+1} = \underbrace{x_k - \alpha_k \nabla f(x_k)}_{\text{gradient descent}} + \underbrace{\beta_k (x_k - x_{k-1})}_{\text{Heavy ball momentum}} .$$

Example: Quadratic cost function $f(x) = \frac{1}{2} x^\top Q x$ with lowest eigenvalue $\lambda_{\min}(Q) = \mu$ and largest eigenvalue $\lambda_{\max}(Q) = L$. $\quad \rightarrow$ condition number $\kappa = \frac{L}{\mu} \geq 1$.

Method	step size	momentum	convergence rate
GD	$\bar{\alpha} = \frac{2}{\mu+L}$	$\beta = 0$	$c = \frac{\kappa-1}{\kappa+1}$
HBM	$\bar{\alpha} = \frac{4}{(\sqrt{\mu}+\sqrt{L})^2}$	$\beta = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$	$c = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$

Accelerated gradient descent method

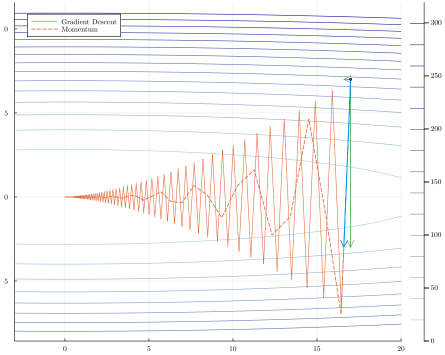


Figure: Illustration of the effect through momentum.

Accelerated gradient descent method

Lower bound on convergence?

Assumption (first order)

The sequence $(x_k)_{k\in\mathbb{N}}$ (generated by some iterative scheme) satisfies the condition

$$x_k \in x_0 + \operatorname{span}\{\nabla f(x_0), \dots, \nabla f(x_{k-1})\}$$

for all $k \geq 1$.

Accelerated gradient descent method

Lower bound on convergence?

Theorem 2.1.13 in Nesterov (2018) - strong convex and smooth

For each $x_0 \in \ell^2(\mathbb{R})$, $\mu, L > 0$ with $\kappa = \frac{L}{\mu} > 1$, there exists a μ-strongly convex and L-smooth function $f : \ell^2(\mathbb{R}) \to \mathbb{R}$ such that every iterative scheme $(x_k)_{k\in\mathbb{N}}$ satisfying Assumption (first order) satisfies a lower bound on the error given by

$$e(x_k) := \|x_k - x_*\|^2 \geq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k} \|x_0 - x_*\|^2,$$

where $x_* \in \ell^2(\mathbb{R})$ denotes the unique global minimum of f.

Upper bound for GD: $e(x_k) := \|x_k - x_*\|^2 \leq \left(\frac{\kappa-1}{\kappa+1}\right)^{2k} \|x_0 - x_*\|^2$

Accelerated gradient descent method

Lower bound on convergence?

Theorem 2.1.7 in Nesterov (2018) - convex and smooth

For every $k \in \mathbb{N}$ with $1 \leq k \leq \frac{1}{2}(d-1)$, $L > 0$ and every $x_0 \in \mathbb{R}^d$ (d denotes the dimension of the domain), there exists a convex and L-smooth function $f : \mathbb{R}^d \to \mathbb{R}$ such that every iterative scheme $(x_k)_{k\in\mathbb{N}}$ satisfying Assumption (first order) satisfies a lower bound on the error given by

$$e(x_k) := f(x_k) - f_* \geq \frac{3L\|x_0 - x_*\|^2}{32(k+1)^2},$$

where $f_* = \min_{x\in\mathbb{R}^d} \ f(x) > -\infty$ exists.

Upper bound for GD: $e(x_k) := f(x_k) - f_* \leq \frac{C}{(k+1)}$

Accelerated gradient descent method

Counter example HBM:
Consider L-smooth and μ-strongly convex function

$$f(x) = \begin{cases} \frac{25}{2}x^2, & x < 1 \\ \frac{1}{2}x^2 + 24x - 12, & x \in [1, 2) \\ \frac{25}{2}x^2 - 24x + 36, & x \geq 2 \end{cases}.$$

Implementation: HBM with $\bar{\alpha} = \frac{4}{(\sqrt{\mu}+\sqrt{L})^2}$, $\beta = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$ and $x_0 = 3.3$.

Accelerated gradient descent method

Counter example HBM:

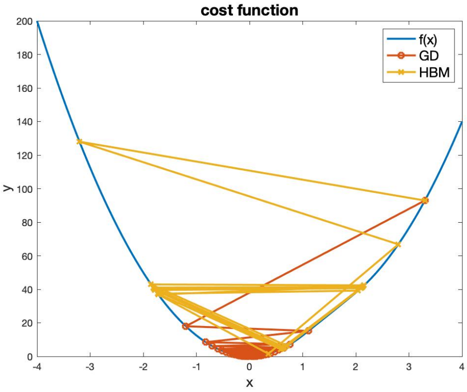


Figure: Evolution of the cost function along the iteration.

Accelerated gradient descent method

Counter example HBM:

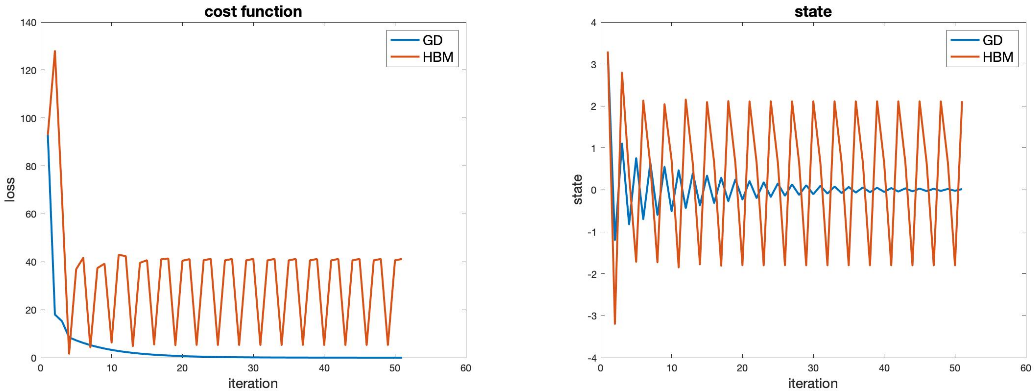


Figure: Evolution of the cost function along the iteration (left) and the state (right).

Accelerated gradient descent method

Nesterov's accelerated gradient descent method:

- cost function $f : \mathbb{R}^d \to \mathbb{R}$,
- step sizes $(\alpha_k)_{k\in\mathbb{N}}$, $\alpha_k > 0$, and momentum parameters $(\beta_k)_{k\in\mathbb{N}}$, $\beta_k \geq 0$,
- initial $q_0, p_0 \in \mathbb{R}^d$.

Iterate:

$$\begin{aligned} p_{k+1} &= q_k - \alpha_k \nabla f(q_k) \\ q_{k+1} &= p_{k+1} + \beta_k (p_{k+1} - p_k) \end{aligned}$$

Accelerated gradient descent method

Nesterov's accelerated gradient descent method: convex case
Written as three variables: (special case $\alpha_k = \gamma_k \tau_k$, $\beta_k = \frac{\tau_{k+1}(1-\tau_k)}{\tau_k}$, $y \mapsto p$, $x \mapsto q$)

$$\begin{aligned} x_k &= \tau_k z_k + (1-\tau_k) y_k, \\ y_{k+1} &= x_k - \alpha_k \nabla f(x_k), \\ z_{k+1} &= z_k - \gamma_k \nabla f(x_k), \end{aligned}$$

Accelerated gradient descent method

Theorem (convex and smooth cost function)

- $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and convex with $\min_{x\in\mathbb{R}^d} f > -\infty$,
- $\alpha_k = \frac{1}{L}$, $A_k > 0$, $\gamma_k = A_{k+1} - A_k \geq 0$ and $\tau_k = \frac{\gamma_k}{A_{k+1}} = \frac{A_{k+1}-A_k}{A_{k+1}} \in (0,1)$,
- initial $(y_0, z_0) \in \mathbb{R}^d \times \mathbb{R}^d$.

Then the increments of $(E_k)_{k\in\mathbb{N}}$ defined as $E_k = \frac{1}{2}\|z_k - x_*\|^2 + A_k(f(y_k) - f(x_*))$ satisfy

$$E_{k+1} - E_k \leq \left(\frac{1}{2}(A_{k+1} - A_k)^2 - \frac{1}{2L}A_{k+1}\right)\|\nabla f(x_k)\|^2$$

for all $k \in \mathbb{N}$. For the particular choice $A_k = \frac{1}{4L}(k+1)k$, $k \geq 1$, and $A_0 = A_1$, we obtain

$$e_k = f(y_k) - f_* \leq \frac{4LE_0}{(k+1)k}, \quad k \geq 1.$$

Accelerated gradient descent method

Nesterov's accelerated gradient descent method: strongly convex case

Written as three variables: (special case $\alpha_k = \frac{1}{L}$, $\beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$, $\tau = \sqrt{\frac{\mu}{L}}$, $y \mapsto p$, $x \mapsto q$)

$$
\begin{aligned}
x_k &= \frac{\tau}{1+\tau} z_k + \frac{1}{1+\tau} y_k \\
y_{k+1} &= x_k - \frac{1}{L} \nabla f(x_k) \\
z_{k+1} &= z_k + \tau(x_k - z_k) - \frac{\tau}{\mu} \nabla f(x_k)
\end{aligned}
$$

Accelerated gradient descent method

Theorem (strongly convex and smooth cost function)

- $f : \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex and L-smooth with $L > \mu$,
- $x_* \in \mathbb{R}^d$ unique global minimum of f,
- $\tau = \sqrt{\frac{\mu}{L}} \in (0, 1)$,
- $(y_0, z_0) \in \mathbb{R}^d \times \mathbb{R}^d$.

Then NAM converges linearly in the sense that

$$e_k := f(y_k) - f(x_*) + \frac{\mu}{2}\|z_k - x_*\|^2 \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \left(f(y_0) - f(x_*) + \frac{\mu}{2}\|z_0 - x_*\|^2\right).$$

Accelerated gradient descent method

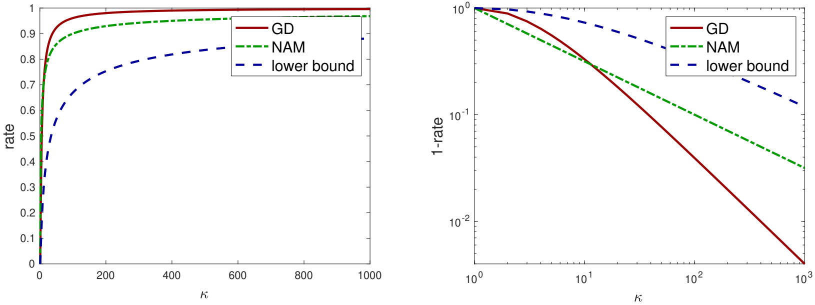


Figure: Illustration of the linear convergence rate depending on the condition number $\kappa = \frac{\mu}{L}$ for GD and NAM. The left plot shows the convergence rate $c^{\mathrm{GD}}(\kappa) = \left(\frac{\kappa-1}{\kappa+1}\right)^2$ and $c^{\mathrm{NAM}}(\kappa) = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}}\right)$, whereas the right plot shows the difference to 1, i.e. $1 - c(\kappa)$, in logarithmic scale.

Ch4: Stochastic approximation in Optimization

Expected and empirical risk

- $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ be $\mathcal{B}(\mathbb{R}^d) \otimes \mathcal{B}(\mathbb{R}^p)/\mathcal{B}(\mathbb{R})$ measurable,
- $Z : \Omega \to \mathbb{R}^p$ random variable with distribution μ_Z, $\mathbb{E}[|f(x, Z)|] < \infty$ for all $x \in \mathbb{R}^d$,
- $Z_1, \ldots, Z_N$ be i.i.d. random variables with $Z_1 \sim \mu_Z$.

Definition

1. *expected risk*:
$$F(x) = \mathbb{E}_{Z\sim\mu}[f(x, Z)] =: \int_{\mathbb{R}^p} f(x, z)\, \mu(\mathrm{d}z), \quad x \in \mathbb{R}^d.$$
2. *empirical risk*:
$$F_N(x) = \frac{1}{N} \sum_{i=1}^{N} f(x, Z^{(i)}).$$

Stochastic gradient descent method

Lemma

Suppose "certain Assumptions" on f and Z are satisfied, then

1. the function $F(x) = \mathbb{E}[f(x, Z)]$ is continuously differentiable,
2. $\nabla f(x, Z)$ is an unbiased estimator of $\nabla_x F(x)$ for every $x \in \mathbb{R}^d$, i.e. it holds true that

$$\nabla_x F(x) = \mathbb{E}[\nabla_x f(x, Z)].$$

Stochastic gradient descent method

Input:

- cost function $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$
- initial random variable $X_0 : \Omega \to \mathbb{R}^d$
- sequence of step sizes $(\alpha_k)_{k\in\mathbb{N}}$, $\alpha_k > 0$ (deterministic or $\mathcal{F}$-adapted)
- sequence of i.i.d. random variables $(Z_k)_{k\in\mathbb{N}}$ with $Z_1 \sim \mu_Z$.

Algorithm: Stochastic gradient descent method (SGD)

- set $k = 0$
- **While** "convergence/stopping criterion not met"
  - approximate the gradient $\nabla_x F(X_k)$ through

$$G_k = \nabla_x f(X_k, Z_{k+1})$$

  - set $X_{k+1} = X_k - \alpha_k G_k$, $k \mapsto k+1$

EndWhile

Stochastic gradient descent method

Input:

- cost function $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$
- initial random variable $X_0 : \Omega \to \mathbb{R}^d$
- sequence of step sizes $(\alpha_k)_{k\in\mathbb{N}}$, $\alpha_k > 0$ (deterministic or $\mathcal{F}$-adapted)
- realization of fixed deterministic data set $\{z^{(i)}\}_{i=1}^N$ with $z^{(i)} \in \mathbb{R}^p$.

Algorithm: SGD with finite data

- set $k = 0$
- **While** "convergence/stopping criterion not met"
  - generate independently $\mathrm{i}_{k+1} \sim \mathcal{U}(\{1, \dots, N\})$
  - approximate the gradient $\nabla_x F_N(X_k)$ through

$$G_k = \nabla_x f(X_k, z^{\mathrm{i}_{k+1}})$$

  - set $X_{k+1} = X_k - \alpha_k G_k$, $k \mapsto k+1$

EndWhile

Convergence analysis of SGD

Decompose the iterative scheme:

$$\begin{aligned} X_{k+1} = X_k - \alpha_k \nabla_x f(X_k, Z_{k+1}) &= X_k - \alpha_k \nabla_x F(X_k) + \alpha_k \left(\nabla_x F(X_k) - \nabla_x f(X_k, Z_{k+1}) \right) \\ &=: X_k - \alpha_k \nabla_x F(X_k) + \alpha_k M_{k+1}. \end{aligned}$$

Factorization:

$$\mathbb{E}[M_{k+1} \mid \mathcal{F}_k] = \mathbb{E}[\nabla_x F(X_k) - \nabla_x f(X_k, Z_{k+1}) \mid \mathcal{F}_k] = 0,$$

where $\mathcal{F}_k = \sigma(X_0,\ Z_m, m \leq k)$

Almost sure convergence of SGD

Robbins & Siegmund

- $(\Omega, \mathcal{A}, \mathcal{F}, \mathbb{P})$ filtered probability space,
- $(Z_k)_{k\in\mathbb{N}}$, $(A_k)_{k\in\mathbb{N}}$, $(B_k)_{k\in\mathbb{N}}$ and $(C_k)_{k\in\mathbb{N}}$ be non-negative and $\mathcal{F}$-adapted,
- $\sum_{k=0}^{\infty} A_k < \infty$ and $\sum_{k=0}^{\infty} B_k < \infty$ almost surely,
- assume

$$\mathbb{E}[Z_{k+1} \mid \mathcal{F}_k] \le Z_k(1 + A_k) + B_k - C_k.$$

Then

1. there exists an almost surely finite random variable Z_∞ such that $Z_k \to Z_\infty$ almost surely for $k \to \infty$,
2. it holds true that $\sum_{k=0}^{\infty} C_k < \infty$ almost surely.

Almost sure convergence of SGD

The SGD iteration satisfies under L-smoothness

$$\mathbb{E}[F(X_{k+1}) - F_* \mid \mathcal{F}_k] \leq (1 + c\frac{L}{2}\alpha_k^2)(F(X_k) - F_*) + c\frac{L}{2}\alpha_k^2 - \alpha_k(1 - \frac{L}{2}\alpha_k)\|\nabla_x F(X_k)\|^2.$$

Theorem (SGD almost sure convergence)

- $F : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and bounded from below by $F_* = \inf_{x \in \mathbb{R}^d} F(x) > -\infty$,
- $\alpha_k > 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty$ and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$ (almost surely),
- suppose that "certain" Assumptions are satisfied,
- X_0 be rv such that $\mathbb{E}[F(X_0)] < \infty$.

Then $(F(X_k))_{k \in \mathbb{N}}$ converges almost surely to some random variable F_∞, almost surely finite, and

$$\lim_{k \to \infty} \|\nabla_x F(X_k)\|^2 = 0, \quad \text{almost surely.}$$

Convergence of SGD (convex)

Theorem (SGD for convex and smooth cost function)

- $F : \mathbb{R}^d \to \mathbb{R}$ be convex and L-smooth, set of global minima of F is non-empty,
- suppose that "certain" Assumptions are satisfied (+ uniform variance bound),
- X_0 be rv such that $\mathbb{E}[|F(X_0)| + \|X_0 - x_*\|^2] < \infty$ for some $x_* \in \arg\min_{x \in \mathbb{R}^d} \; F(x)$,
- $\alpha_k \in (0, \frac{1}{L}]$, deterministic and decreasing.

Then for $\bar{X}_N := \sum_{k=0}^{N-1} w_k^N X_{k+1}, \; w_k^N := \frac{\alpha_k}{\sum_{j=0}^{N-1} \alpha_j}, \; N \geq 2$, it holds true that

$$\mathbb{E}[F(\bar{X}_N) - F(x_*)] \leq \frac{\mathbb{E}[\|X_0 - x_*\|^2]}{2\sum_{j=0}^{N-1} \alpha_j} + \frac{c(1+\alpha_0 L)\sum_{k=0}^{N-1} \alpha_k^2}{2\sum_{j=0}^{N-1} \alpha_j}.$$

$$\alpha_k := \frac{1}{L\sqrt{k+1}} \quad \Longrightarrow \quad \mathbb{E}[F(\bar{X}_N) - F(x_*)] \in \mathcal{O}\left(\frac{\log(N)}{\sqrt{N}}\right).$$

Convergence of SGD (strongly convex)

Theorem (SGD for strongly convex and smooth cost function)

- $F : \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex and L-smooth,
- suppose that "certain" Assumptions are satisfied (+ uniform variance bound),
- X_0 be rv such that $\mathbb{E}[|F(X_0)| + \|X_0 - x_*\|^2] < \infty$, $x_* \in \mathbb{R}^d$ global minimum of F,
- $\alpha_k \in (0, \frac{1}{L}]$, deterministic.

Then for all $k \geq 0$ it holds true that

$$\mathbb{E}[\|X_{k+1} - x_*\|^2] \leq (1 - \alpha_k \mu)\mathbb{E}[\|X_k - x_*\|^2] + c\alpha_k^2$$

$$\alpha_k := \frac{\tau}{\mu(k+s)} \quad \Longrightarrow \quad \mathbb{E}[\|X_k - x_*\|^2] \in \mathcal{O}\left(\frac{1}{k+s}\right).$$

Variance reduction for SGD

Assumption	error bound
convex	$\frac{C_1}{\sqrt{k}} + \mathrm{var} \cdot \frac{\log(k)}{\sqrt{k}}$
strong convex	$(1 - \alpha_k \mu) e_k + \mathrm{var} \cdot \alpha_k^2$
PL-condition	$(1 - \alpha_k r) e_k + \mathrm{var} \cdot \alpha_k^2$

Dynamical sampling

Input:

- cost function $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$
- initial random variable $X_0 : \Omega \to \mathbb{R}^d$
- sequence of step sizes $(\alpha_k)_{k\in\mathbb{N}}$, $\alpha_k > 0$ (deterministic or $\mathcal{F}$-adapted)
- sequence of batch sizes $(B_k)_{k\in\mathbb{N}}$
- sequence of i.i.d. random variables $(Z_k^{(m)})_{k\in\mathbb{N},\ m=1,\dots,B_{k-1}}$ with $Z_1^{(1)} \sim \mu_Z$.

Algorithm: SGD with dynamical sampling

- set $k = 0$
- **While** "convergence/stopping criterion not met"
  - approximate the gradient $\nabla_x F(X_k)$ through

$$G_k = \frac{1}{B_k} \sum_{m=1}^{B_k} \nabla_x f(X_k, Z_{k+1}^{(m)})$$

  - set $X_{k+1} = X_k - \alpha_k G_k$, $k \mapsto k+1$

EndWhile

Dynamical sampling

Optimal dynamical batch-size: $B_j = \varepsilon^{-1} 2c\bar{\alpha}^2 \left(\frac{1-\rho^{\frac{K}{2}}}{1-\rho^{\frac{1}{2}}} \right) \rho^{\frac{K-1-j}{2}}$, with computational cost

$$\sum_{j=0}^{K-1} B_j = \varepsilon^{-1} 2c\bar{\alpha}^2 \left(\frac{1-\rho^{\frac{K}{2}}}{1-\rho^{\frac{1}{2}}} \right) \sum_{j=0}^{K-1} \rho^{\frac{K-1-j}{2}} \simeq \varepsilon^{-1},$$

Fixed batch-size: $\bar{B} \geq \varepsilon^{-1} 2c\bar{\alpha}^2 (1-\rho)^{-1} \simeq \varepsilon^{-1}$, with computational cost

$$\sum_{j=0}^{K-1} B_j = K \cdot \bar{B} \simeq |\log(\varepsilon^{-1})| \varepsilon^{-1} .$$

Conclusion

Outlook

- adaptive step sizes (Adagrad, Adadelta,...)
- incorporation of momentum into SGD
- adaptive moment estimation (ADAM) ➜ combining everything
- many more variants of SGD...
- other (heuristic) algorithms (simulated annealing, particle swarm optimization,...)
- Application to specific machine learning models (Regression, support vector machines, neural networks, GP's)

Additional information

- Seminar *Stochastik*
- Master thesis possible on related topics