Optimization for data science

Smooth optimization: Gradient descent

R. Flamary

Master Data Science, Institut Polytechnique de Paris

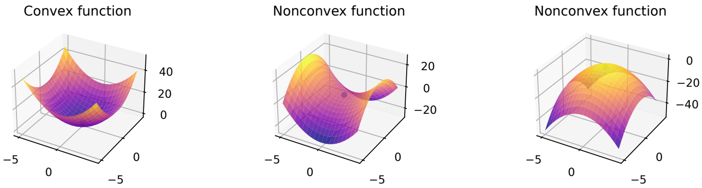
September 17, 2024

Full course overview

1. **Introduction to optimization for data science**
   1.1 ML optimization problems and linear algebra recap
   1.2 Optimization problems and their properties (Convexity, smoothness)
2. **Smooth optimization : Gradient descent**
   2.1 First order algorithms, convergence for smooth and strongly convex functions
3. **Smooth Optimization : Quadratic problems**
   3.1 Solvers for quadratic problems, conjugate gradient
   3.2 Linesearch methods
4. **Non-smooth Optimization : Proximal methods**
   4.1 Proximal operator and proximal algorithms
   4.2 Lab 1: Lasso and group Lasso
5. **Stochastic Gradient Descent**
   5.1 SGD and variance reduction techniques
   5.2 Lab 2: SGD for Logistic regression
6. **Standard formulation of constrained optimization problems**
   6.1 LP, QP and Mixed Integer Programming
7. **Coordinate descent**
   7.1 Algorithms and Labs
8. **Newton and quasi-newton methods**
   8.1 Second order methods and Labs
9. **Beyond convex optimization**
   9.1 Nonconvex reg., Frank-Wolfe, DC programming, autodiff

Current course overview

Smooth Optimization problem



Optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; F(\mathbf{x}), \tag{1}$$

- F is L-smooth (at least differentiable).
- When F is convex $\mathbf{x}^\star$ is a solution of the problem if

$$\nabla_{\mathbf{x}} F(\mathbf{x}^\star) = \mathbf{0}$$

- When F is non convex $\mathbf{x}^\star$ is a local minimizer of the problem if

$$\nabla_{\mathbf{x}} F(\mathbf{x}^\star) = \mathbf{0} \qquad \text{and} \qquad \nabla^2_{\mathbf{x}} F(\mathbf{x}^\star) \succeq 0$$

How to solve optimization problems?

- Solving the problem analytically : $\nabla F(\mathbf{x}^\star) = 0$
- Search for a solution numerically : iterative optimization algorithms

Iterative optimization algorithms

$$\min_{\mathbf{x}\in\mathbb{R}^n} \; F(\mathbf{x}),$$

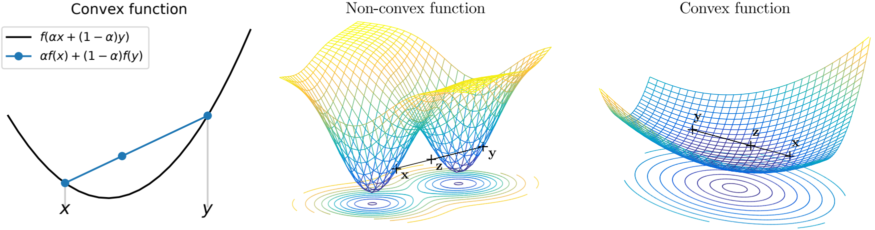
Iterative algorithms

- Principle : start from an initial point $\mathbf{x}^{(0)}$ and iterate to make it better.
- Gradient descent (and variants) when available, proximal methods.
- Black box optimization (a.k.a derivative free optimization) :
  - Genetic, random search, simulated annealing [Gen and Cheng, 1999].
  - Particle swarm optimization, etc [Kennedy and Eberhart, 1995].
  - Nelder-Mead simplex [Nelder and Mead, 1965].

How to choose?

- No free lunch theorem [Wolpert and Macready, 1997] :
  No algorithm is better than the others for all problems.
- But on can use the properties of the problem to choose the algorithm: specialize!

Assumption 1 : Convexity



Convex function (recap)

- Function F is **convex** if it lies below its chords, that is $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$F(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha F(\mathbf{x}) + (1-\alpha)F(\mathbf{y}), \text{ with } 0 \leq \alpha \leq 1. \quad (2)$$

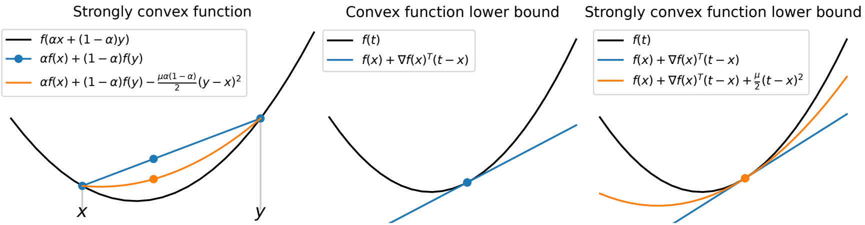
- F a differentiable function is **convex** if and only if

$$F(\mathbf{y}) \geq F(\mathbf{x}) + \nabla F(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}), \quad \forall \mathbf{y}, \mathbf{x} \in \text{dom} F \quad (3)$$

- For $\mathcal{C} = \mathbb{R}^n$, if $\mathbf{x}$ if a global minimum if and only if $\nabla_{\mathbf{x}} F(\mathbf{x}) = \mathbf{0}$.
- F is μ-**strongly convex** with $\mu > 0$ if it satisfies $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $0 \leq \alpha \leq 1$

$$F(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha F(\mathbf{x}) + (1-\alpha)F(\mathbf{y}) - \frac{\mu}{2}\alpha(1-\alpha)\|\mathbf{x} - \mathbf{y}\|^2, \quad (4)$$

Assumption 1 : Convexity



Convex function (recap)

- Function F is **convex** if it lies below its chords, that is $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$F(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha F(\mathbf{x}) + (1-\alpha)F(\mathbf{y}), \text{ with } 0 \leq \alpha \leq 1. \tag{2}$$

- F a differentiable function is **convex** if and only if

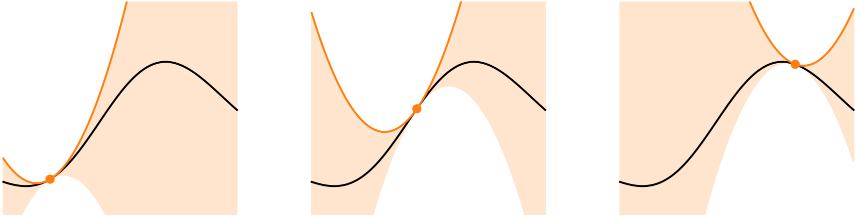
$$F(\mathbf{y}) \geq F(\mathbf{x}) + \nabla F(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}), \quad \forall \mathbf{y}, \mathbf{x} \in \text{dom} F \tag{3}$$

- For $\mathcal{C} = \mathbb{R}^n$, if $\mathbf{x}$ if a global minimum if and only if $\nabla_{\mathbf{x}} F(\mathbf{x}) = \mathbf{0}$.
- F is μ-**strongly convex** with $\mu > 0$ if it satisfies $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $0 \leq \alpha \leq 1$

$$F(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha F(\mathbf{x}) + (1-\alpha)F(\mathbf{y}) - \frac{\mu}{2}\alpha(1-\alpha)\|\mathbf{x} - \mathbf{y}\|^2, \tag{4}$$

Assumption 2 : smoothness

L-smooth function



L-smooth function (recap)

- Function F is **gradient Lipschitz**, also called L-**smooth**, if $\forall \mathbf{x}, \mathbf{y} \in \mathcal{C}^2$

$$\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \tag{5}$$

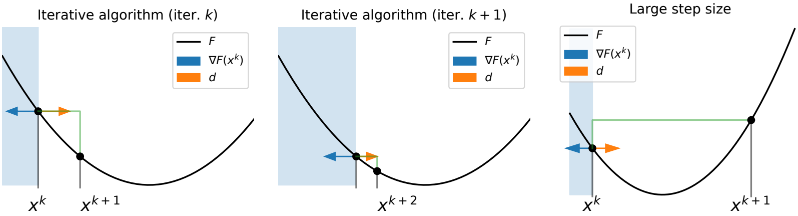
- If F is L-**smooth**, then the following inequality holds

$$F(\mathbf{x}) \leq F(\mathbf{y}) + \nabla F(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2 \tag{6}$$

- If F is L-**smooth**, then the following inequality holds

$$\nabla^2_{\mathbf{x}} F(\mathbf{x}) \preceq L\mathbf{I} \quad (\lambda_{\text{max}}(\nabla^2_{\mathbf{x}} F(\mathbf{x})) \leq L) \tag{7}$$

Descent algorithm for smooth optimization

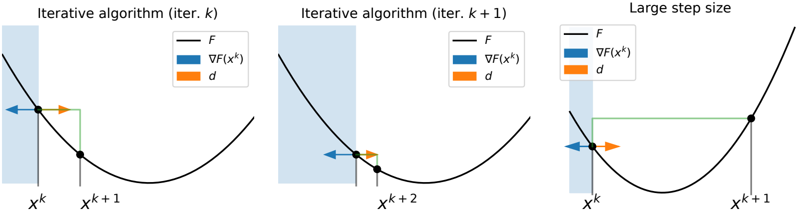


General iterative algorithm

Initialize $\mathbf{x}^{(0)}$
for $k = 0, 1, 2, \dots$ **do**
  $\mathbf{d}^{(k)} \leftarrow$ Compute descent direction from $\mathbf{x}^{(k)}$
  $\rho^{(k)} \leftarrow$ Choose stepsize
  $\mathbf{x}^{(k+1)} \leftarrow \mathbf{x}^{(k)} + \rho^{(k)}\mathbf{d}^{(k)}$
end for

- $\mathbf{x}^{(k)} \in \mathbb{R}^n$ is the current iterate.
- $\mathbf{d}^{(k)} \in \mathbb{R}^n$ is a descent direction if $\nabla F(\mathbf{x}^{(k)})^T\mathbf{d}^{(k)} < 0$.
- For a step small enough, each iteration decreases the cost : $F(\mathbf{x}^{(k+1)}) \leq F(\mathbf{x}^{(k)})$
- Stopping conditions: max number of iterations or small gradient $\|\nabla F(\mathbf{x}^k)\|$.

Gradient Descent (GD) algorithm

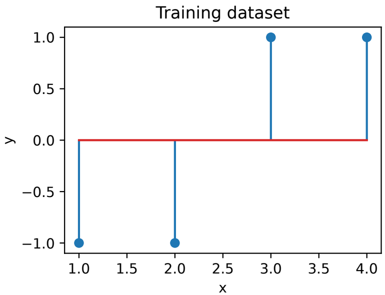


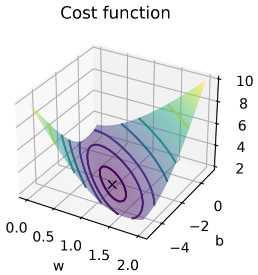
Gradient descent algorithm (steepest descent)

Initialize $\mathbf{x}^{(0)}$
for $k = 0, 1, 2, \dots$ **do**
$\quad \mathbf{d}^{(k)} \leftarrow -\nabla F(\mathbf{x}^{(k)})$
$\quad \rho^{(k)} \leftarrow$ Choose stepsize
$\quad \mathbf{x}^{(k+1)} \leftarrow \mathbf{x}^{(k)} + \rho^{(k)}\mathbf{d}^{(k)}$
end for

- Iterative algorithm with descent direction $\mathbf{d} = -\nabla F(\mathbf{x})$.
- $-\nabla F(\mathbf{x})$ is called the steepest descent direction.
- Equivalent to iterative algorithm above in 1D.
- In this course we study the constant step case $\rho^{(k)} = \rho$.

Example optimization problem



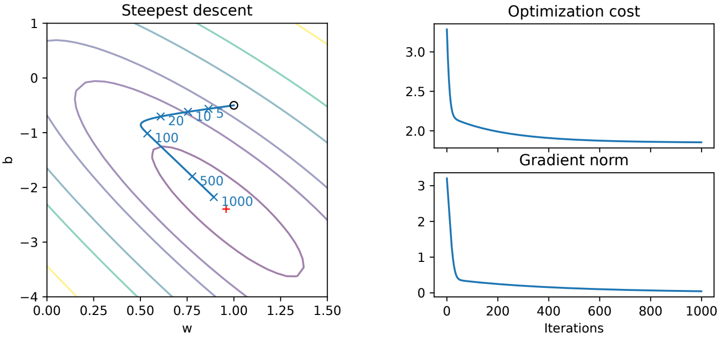


1D Logistic regression

$$\min_{w,b} \quad \sum_{i=1}^{n} \log(1 + \exp(-y_i(wx_i + b))) + \lambda \frac{w^2}{2}$$

- Linear prediction model : $f(x) = wx + b$
- Training data (x_i, y_i) : $(1, -1), (2, -1), (3, 1), (4, 1)$.
- Problem solution for $\lambda = 1$: $\mathbf{x}^* = [w^\star, b^\star] = [0.96, -2.40]$
- Initialization : $\mathbf{x}^{(0)} = [1, -0.5]$.
- Complexity : Cost and gradient both $O(nd)$

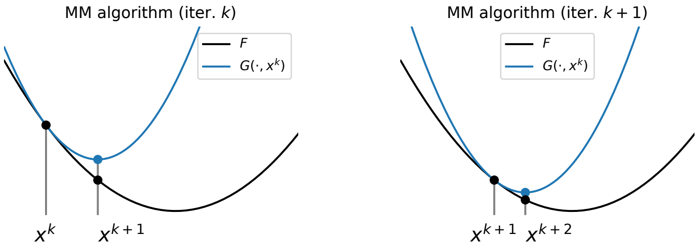
Example of steepest descent



Discussion

- Steepest descent with fixed step $\rho^{(k)} = 0.1$
- Slow convergence around the solution (small gradients).
- After 1000 iterations, still not converged.
- Complexity $\mathcal{O}(nd)$ per iteration.

Majorization Minimization (MM) algorithm



Principle

- Iterative algorithm that minimizes a surrogate function.
- Let F be a function to minimize and G a majorization $F(\mathbf{x}) \leq G(\mathbf{x}, \mathbf{y}) \quad \forall \mathbf{x}, \mathbf{y}$.
- MM iteration :

$$\mathbf{x}^{(k+1)} = \underset{\mathbf{x}}{\text{argmin}} \quad G(\mathbf{x}, \mathbf{x}^{(k)}) \tag{8}$$

- The MM algorithm is guaranteed to decrease the cost function at each iteration.
- Most efficient when G is close to F, but simple to compute and optimize.
- References : [Hunter and Lange, 2004, Sun et al., 2016].

Majorization Minimization for smooth functions

Majorization of L-smooth functions

If F is L-smooth, then the following majorization holds:

$$F(\mathbf{x}) \leq G(\mathbf{x}, \mathbf{y}) = F(\mathbf{y}) + \nabla F(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 \tag{9}$$

Solving the MM iteration with quadratic upper bound

$$x^{(k+1)} = \underset{\mathbf{x}}{\text{argmin}} \quad F(\mathbf{x}^{(k)}) + \nabla F(\mathbf{x}^{(k)})^\top (\mathbf{x} - \mathbf{x}^{(k)}) + \frac{L}{2} \|\mathbf{x} - \mathbf{x}^{(k)}\|^2 \tag{10}$$

- The MM iteration is a quadratic problem that can be solved analytically.
- The solution is given by:

Majorization Minimization for smooth functions

Majorization of L-smooth functions

If F is L-smooth, then the following majorization holds:

$$F(\mathbf{x}) \leq G(\mathbf{x}, \mathbf{y}) = F(\mathbf{y}) + \nabla F(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 \tag{9}$$

Solving the MM iteration with quadratic upper bound

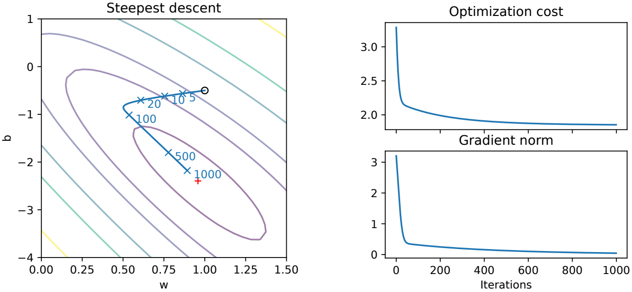
$$x^{(k+1)} = \underset{\mathbf{x}}{\text{argmin}} \quad F(\mathbf{x}^{(k)}) + \nabla F(\mathbf{x}^{(k)})^\top (\mathbf{x} - \mathbf{x}^{(k)}) + \frac{L}{2} \|\mathbf{x} - \mathbf{x}^{(k)}\|^2 \tag{10}$$

- The MM iteration is a quadratic problem that can be solved analytically.
- The solution is given by:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \frac{1}{L} \nabla F(\mathbf{x}^{(k)}) \tag{11}$$

- This is exactly the update of the gradient descent with step $\rho = \frac{1}{L}$.

Convergence of gradient descent



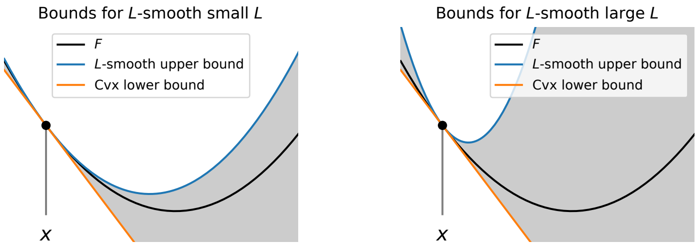
Questions

- Does Gradient descent converges to an optimal point ?
- At which speed is the minimum reached?
- How to choose the stepsize $\rho^{(k)}$?

Theoretical convergence and convergence speed

- Fixed steps $\rho^{(k)} = \rho$?
- Smooth and strongly convex functions ?
- Acceleration techniques ?
- Adaptive steps $\rho^{(k)}$ (linesearch, next course) ?

Convergence for smooth functions



Convergence of gradient descent for L-smooth functions

If function F is convex and differentiable and its gradient has a Lipschitz constant L, then the gradient descent with fixed step $\rho^{(k)} = \rho \leq \frac{1}{L}$ converges to a solution $\mathbf{x}^\star$ of the optimization problem with the following speed:

$$F(\mathbf{x}^{(k)}) - F(\mathbf{x}^\star) \leq \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^\star\|^2}{2\rho k} \quad (12)$$

- Best for $\rho = \frac{1}{L}$ that is the largest gradient that ensures decrease of the cost.
- We say the the gradient descent has a convergence $O(\frac{1}{k})$.
- In order to reach a precision ϵ one needs $O(\frac{1}{\epsilon})$ iterations.
- We prove this result in the next slides [1].

[1] See also : `https://www.stat.cmu.edu/~ryantibs/convexopt-F13/scribes/lec6.pdf`

Convergence proof (convex L-smooth)

Step 1 : Descent VS gradient norm Lemma

$$F(\mathbf{x}^{(k+1)}) \leq F(\mathbf{x}^{(k)}) - \frac{\rho}{2}\|\nabla F(\mathbf{x}^{(k)})\|^2 \tag{13}$$

Value decreases at each iteration for $\rho \leq \frac{1}{L}$.

Proof.

$$F(\mathbf{x}^{(k+1)}) \underset{2}{\leq} F(\mathbf{x}^{(k)}) + \nabla F(\mathbf{x}^{(k)})^T(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) + \frac{L}{2}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|^2$$

$$\underset{3}{=}$$

□

[2] Convexity upper bound w.r.t. $\mathbf{x}^{(k)}$

[3] Inject gradient step $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \rho\nabla F(\mathbf{x}^{(k)})$

[4] For $\rho \leq \frac{1}{L}$, $-(2 - \rho L) \leq -1$

Convergence proof (convex L-smooth)

Step 1 : Descent VS gradient norm Lemma

$$F(\mathbf{x}^{(k+1)}) \leq F(\mathbf{x}^{(k)}) - \frac{\rho}{2}\|\nabla F(\mathbf{x}^{(k)})\|^2 \tag{13}$$

Value decreases at each iteration for $\rho \leq \frac{1}{L}$.

Proof.

$$F(\mathbf{x}^{(k+1)}) \underset{2}{\leq} F(\mathbf{x}^{(k)}) + \nabla F(\mathbf{x}^{(k)})^T(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) + \frac{L}{2}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|^2$$

$$\underset{3}{=} F(\mathbf{x}^{(k)}) + \nabla F(\mathbf{x}^{(k)})^T(-\rho\nabla F(\mathbf{x}^{(k)})) + \frac{L}{2}\| - \rho\nabla F(\mathbf{x}^{(k)})\|^2$$

$$=$$

□

[2] Convexity upper bound w.r.t. $\mathbf{x}^{(k)}$

[3] Inject gradient step $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \rho\nabla F(\mathbf{x}^{(k)})$

[4] For $\rho \leq \frac{1}{L}$, $-(2-\rho L) \leq -1$

Convergence proof (convex L-smooth)

Step 1 : Descent VS gradient norm Lemma

$$F(\mathbf{x}^{(k+1)}) \leq F(\mathbf{x}^{(k)}) - \frac{\rho}{2}\|\nabla F(\mathbf{x}^{(k)})\|^2 \tag{13}$$

Value decreases at each iteration for $\rho \leq \frac{1}{L}$.

Proof.

$$\begin{aligned}
F(\mathbf{x}^{(k+1)}) &\underset{2}{\leq} F(\mathbf{x}^{(k)}) + \nabla F(\mathbf{x}^{(k)})^T(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) + \frac{L}{2}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|^2 \\
&\underset{3}{=} F(\mathbf{x}^{(k)}) + \nabla F(\mathbf{x}^{(k)})^T(-\rho\nabla F(\mathbf{x}^{(k)})) + \frac{L}{2}\| - \rho\nabla F(\mathbf{x}^{(k)})\|^2 \\
&= F(\mathbf{x}^{(k)}) - \rho\|\nabla F(\mathbf{x}^{(k)})\|^2 + \frac{L\rho^2}{2}\|\nabla F(\mathbf{x}^{(k)})\|^2 \\
&=
\end{aligned}$$

□

[2] Convexity upper bound w.r.t. $\mathbf{x}^{(k)}$

[3] Inject gradient step $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \rho\nabla F(\mathbf{x}^{(k)})$

[4] For $\rho \leq \frac{1}{L}$, $-(2 - \rho L) \leq -1$

Convergence proof (convex L-smooth)

Step 1 : Descent VS gradient norm Lemma

$$F(\mathbf{x}^{(k+1)}) \leq F(\mathbf{x}^{(k)}) - \frac{\rho}{2}\|\nabla F(\mathbf{x}^{(k)})\|^2 \quad (13)$$

Value decreases at each iteration for $\rho \leq \frac{1}{L}$.

Proof.

$$\begin{aligned}
F(\mathbf{x}^{(k+1)}) &\underset{2}{\leq} F(\mathbf{x}^{(k)}) + \nabla F(\mathbf{x}^{(k)})^T(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) + \frac{L}{2}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|^2 \\
&\underset{3}{=} F(\mathbf{x}^{(k)}) + \nabla F(\mathbf{x}^{(k)})^T(-\rho\nabla F(\mathbf{x}^{(k)})) + \frac{L}{2}\| - \rho\nabla F(\mathbf{x}^{(k)})\|^2 \\
&= F(\mathbf{x}^{(k)}) - \rho\|\nabla F(\mathbf{x}^{(k)})\|^2 + \frac{L\rho^2}{2}\|\nabla F(\mathbf{x}^{(k)})\|^2 \\
&= F(\mathbf{x}^{(k)}) - \frac{\rho}{2}\|\nabla F(\mathbf{x}^{(k)})\|^2(2 - \rho L) \\
&\underset{4}{\leq} F(\mathbf{x}^{(k)}) - \frac{\rho}{2}\|\nabla F(\mathbf{x}^{(k)})\|^2
\end{aligned}$$

□

[2] Convexity upper bound w.r.t. $\mathbf{x}^{(k)}$

[3] Inject gradient step $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \rho\nabla F(\mathbf{x}^{(k)})$

[4] For $\rho \leq \frac{1}{L}$, $-(2 - \rho L) \leq -1$

Convergence proof (convex L-smooth)

Step 2 : Objective w.r.t. optimal value

$$F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^\star) \leq \frac{1}{2\rho}(\|\mathbf{x}^k - \mathbf{x}^\star\|^2 - \|\mathbf{x}^{k+1} - \mathbf{x}^\star\|^2) \tag{14}$$

Proof.

Using convexity one has: $F(\mathbf{x}) \leq F(\mathbf{x}^\star) + \nabla F(\mathbf{x})^\top (\mathbf{x} - \mathbf{x}^\star)$ so from (13):

$$F(\mathbf{x}^{(k+1)}) \leq$$

Convergence proof (convex L-smooth)

Step 2 : Objective w.r.t. optimal value

$$F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^\star) \leq \frac{1}{2\rho}(\|\mathbf{x}^k - \mathbf{x}^\star\|^2 - \|\mathbf{x}^{k+1} - \mathbf{x}^\star\|^2) \tag{14}$$

Proof.

Using convexity one has: $F(\mathbf{x}) \leq F(\mathbf{x}^\star) + \nabla F(\mathbf{x})^\top(\mathbf{x} - \mathbf{x}^\star)$ so from (13):

$$\begin{aligned} F(\mathbf{x}^{(k+1)}) &\leq F(\mathbf{x}^{(k)}) - \frac{\rho}{2}\|\nabla F(\mathbf{x}^{(k)})\|^2 \\ &\leq \end{aligned}$$

Convergence proof (convex L-smooth)

Step 2 : Objective w.r.t. optimal value

$$F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^\star) \leq \frac{1}{2\rho}(\|\mathbf{x}^k - \mathbf{x}^\star\|^2 - \|\mathbf{x}^{k+1} - \mathbf{x}^\star\|^2) \tag{14}$$

Proof.

Using convexity one has: $F(\mathbf{x}) \leq F(\mathbf{x}^\star) + \nabla F(\mathbf{x})^\top(\mathbf{x} - \mathbf{x}^\star)$ so from (13):

$$\begin{aligned} F(\mathbf{x}^{(k+1)}) &\leq F(\mathbf{x}^{(k)}) - \frac{\rho}{2}\|\nabla F(\mathbf{x}^{(k)})\|^2 \\ &\leq F(\mathbf{x}^\star) + \nabla F(\mathbf{x}^{(k)})^\top(\mathbf{x}^{(k)} - \mathbf{x}^\star) - \frac{\rho}{2}\|\nabla F(\mathbf{x}^{(k)})\|^2 \end{aligned}$$

$F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^\star) \leq$

Convergence proof (convex L-smooth)

Step 2 : Objective w.r.t. optimal value

$$F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^\star) \leq \frac{1}{2\rho}(\|\mathbf{x}^k - \mathbf{x}^\star\|^2 - \|\mathbf{x}^{k+1} - \mathbf{x}^\star\|^2) \tag{14}$$

Proof.

Using convexity one has: $F(\mathbf{x}) \leq F(\mathbf{x}^\star) + \nabla F(\mathbf{x})^\top(\mathbf{x} - \mathbf{x}^\star)$ so from (13):

$$\begin{aligned}
F(\mathbf{x}^{(k+1)}) &\leq F(\mathbf{x}^{(k)}) - \frac{\rho}{2}\|\nabla F(\mathbf{x}^{(k)})\|^2 \\
&\leq F(\mathbf{x}^\star) + \nabla F(\mathbf{x}^{(k)})^\top(\mathbf{x}^{(k)} - \mathbf{x}^\star) - \frac{\rho}{2}\|\nabla F(\mathbf{x}^{(k)})\|^2 \\
F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^\star) &\leq \nabla F(\mathbf{x}^{(k)})^\top(\mathbf{x}^{(k)} - \mathbf{x}^\star) - \frac{\rho}{2}\|\nabla F(\mathbf{x}^{(k)})\|^2 \\
&\leq
\end{aligned}$$

Convergence proof (convex L-smooth)

Step 2 : Objective w.r.t. optimal value

$$F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^\star) \leq \frac{1}{2\rho}(\|\mathbf{x}^k - \mathbf{x}^\star\|^2 - \|\mathbf{x}^{k+1} - \mathbf{x}^\star\|^2) \tag{14}$$

Proof.

Using convexity one has: $F(\mathbf{x}) \leq F(\mathbf{x}^\star) + \nabla F(\mathbf{x})^\top(\mathbf{x} - \mathbf{x}^\star)$ so from (13):

$$\begin{aligned}
F(\mathbf{x}^{(k+1)}) &\leq F(\mathbf{x}^{(k)}) - \frac{\rho}{2}\|\nabla F(\mathbf{x}^{(k)})\|^2 \\
&\leq F(\mathbf{x}^\star) + \nabla F(\mathbf{x}^{(k)})^\top(\mathbf{x}^{(k)} - \mathbf{x}^\star) - \frac{\rho}{2}\|\nabla F(\mathbf{x}^{(k)})\|^2 \\
F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^\star) &\leq \nabla F(\mathbf{x}^{(k)})^\top(\mathbf{x}^{(k)} - \mathbf{x}^\star) - \frac{\rho}{2}\|\nabla F(\mathbf{x}^{(k)})\|^2 \\
&\leq \frac{1}{2\rho}\Big(2\rho\nabla F(\mathbf{x}^{(k)})^\top(\mathbf{x}^{(k)} - \mathbf{x}^\star) - \rho^2\|\nabla F(\mathbf{x}^{(k)})\|^2 - \|\mathbf{x}^{(k)} - \mathbf{x}^\star\|^2 \\
&\qquad + \|\mathbf{x}^{(k)} - \mathbf{x}^\star\|^2\Big) \\
&\underset{5}{\leq}
\end{aligned}$$

Convergence proof (convex L-smooth)

Step 2 : Objective w.r.t. optimal value

$$F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^\star) \leq \frac{1}{2\rho}(\|\mathbf{x}^k - \mathbf{x}^\star\|^2 - \|\mathbf{x}^{k+1} - \mathbf{x}^\star\|^2) \tag{14}$$

Proof.

Using convexity one has: $F(\mathbf{x}) \leq F(\mathbf{x}^\star) + \nabla F(\mathbf{x})^\top(\mathbf{x} - \mathbf{x}^\star)$ so from (13):

$$\begin{aligned}
F(\mathbf{x}^{(k+1)}) &\leq F(\mathbf{x}^{(k)}) - \frac{\rho}{2}\|\nabla F(\mathbf{x}^{(k)})\|^2 \\
&\leq F(\mathbf{x}^\star) + \nabla F(\mathbf{x}^{(k)})^\top(\mathbf{x}^{(k)} - \mathbf{x}^\star) - \frac{\rho}{2}\|\nabla F(\mathbf{x}^{(k)})\|^2 \\
F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^\star) &\leq \nabla F(\mathbf{x}^{(k)})^\top(\mathbf{x}^{(k)} - \mathbf{x}^\star) - \frac{\rho}{2}\|\nabla F(\mathbf{x}^{(k)})\|^2 \\
&\leq \frac{1}{2\rho}\Big(2\rho\nabla F(\mathbf{x}^{(k)})^\top(\mathbf{x}^{(k)} - \mathbf{x}^\star) - \rho^2\|\nabla F(\mathbf{x}^{(k)})\|^2 - \|\mathbf{x}^{(k)} - \mathbf{x}^\star\|^2 \\
&\qquad + \|\mathbf{x}^{(k)} - \mathbf{x}^\star\|^2\Big) \\
&\underset{5}{\leq} \frac{1}{2\rho}\Big(-\|\mathbf{x}^{(k)} - \rho\nabla F(\mathbf{x}^{(k)}) - \mathbf{x}^\star\|^2 + \|\mathbf{x}^{(k)} - \mathbf{x}^\star\|^2\Big) \\
&= \frac{1}{2\rho}(\|\mathbf{x}^k - \mathbf{x}^\star\|^2 - \|\mathbf{x}^{k+1} - \mathbf{x}^\star\|^2)
\end{aligned}$$

[5] Factorization of $\|\mathbf{x}^{(k)} - \rho\nabla F(\mathbf{x}^{(k)}) - \mathbf{x}^\star\|^2$

Convergence proof (convex L-smooth)

Step 3 : Putting all iterations together

$$F(\mathbf{x}^{(k)}) - F(\mathbf{x}^\star) \leq \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^\star\|^2}{2\rho k}$$

Proof.

$$F(\mathbf{x}^{(k)}) - F(\mathbf{x}^\star) = \frac{1}{k}\sum_{i=1}^{k} F(\mathbf{x}^{(k)}) - F(\mathbf{x}^\star)$$

$$\underset{6}{\leq}$$

□

[6] Descent Lemma (13)
[7] Inject Eq. (14)
[8] Summation of telescopic series

Convergence proof (convex L-smooth)

Step 3 : Putting all iterations together

$$F(\mathbf{x}^{(k)}) - F(\mathbf{x}^\star) \leq \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^\star\|^2}{2\rho k}$$

Proof.

$$\begin{aligned} F(\mathbf{x}^{(k)}) - F(\mathbf{x}^\star) &= \frac{1}{k}\sum_{i=1}^{k} F(\mathbf{x}^{(k)}) - F(\mathbf{x}^\star) \\ &\underset{6}{\leq} \frac{1}{k}\sum_{i=1}^{k} F(\mathbf{x}^{(i)}) - F(\mathbf{x}^\star) \\ &\underset{7}{\leq} \end{aligned}$$

□

[6] Descent Lemma (13)
[7] Inject Eq. (14)
[8] Summation of telescopic series

Convergence proof (convex L-smooth)

Step 3 : Putting all iterations together

$$F(\mathbf{x}^{(k)}) - F(\mathbf{x}^\star) \leq \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^\star\|^2}{2\rho k}$$

Proof.

$$\begin{aligned}
F(\mathbf{x}^{(k)}) - F(\mathbf{x}^\star) &= \frac{1}{k}\sum_{i=1}^{k} F(\mathbf{x}^{(k)}) - F(\mathbf{x}^\star) \\
&\underset{6}{\leq} \frac{1}{k}\sum_{i=1}^{k} F(\mathbf{x}^{(i)}) - F(\mathbf{x}^\star) \\
&\underset{7}{\leq} \frac{1}{2\rho k}\sum_{i=1}^{k} \|\mathbf{x}^{k-1} - \mathbf{x}^\star\|^2 - \|\mathbf{x}^{k} - \mathbf{x}^\star\|^2 \\
&\underset{8}{=}
\end{aligned}$$

□

[6] Descent Lemma (13)
[7] Inject Eq. (14)
[8] Summation of telescopic series

Convergence proof (convex L-smooth)

Step 3 : Putting all iterations together

$$F(\mathbf{x}^{(k)}) - F(\mathbf{x}^\star) \leq \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^\star\|^2}{2\rho k}$$

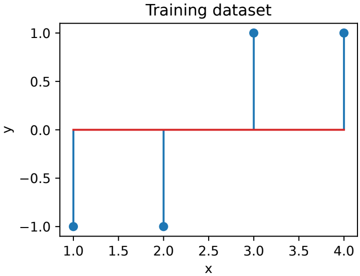
Proof.

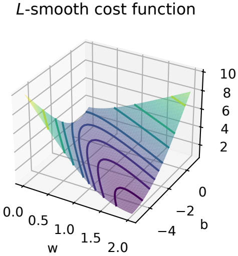
$$\begin{aligned}
F(\mathbf{x}^{(k)}) - F(\mathbf{x}^\star) &= \frac{1}{k}\sum_{i=1}^{k} F(\mathbf{x}^{(k)}) - F(\mathbf{x}^\star) \\
&\underset{6}{\leq} \frac{1}{k}\sum_{i=1}^{k} F(\mathbf{x}^{(i)}) - F(\mathbf{x}^\star) \\
&\underset{7}{\leq} \frac{1}{2\rho k}\sum_{i=1}^{k} \|\mathbf{x}^{k-1} - \mathbf{x}^\star\|^2 - \|\mathbf{x}^{k} - \mathbf{x}^\star\|^2 \\
&\underset{8}{=} \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^\star\|^2 - \|\mathbf{x}^{(k)} - \mathbf{x}^\star\|^2}{2\rho k} \\
&\leq \frac{\|\mathbf{x}^{(0)} - \mathbf{x}^\star\|^2}{2\rho k}
\end{aligned}$$

□

[6] Descent Lemma (13)
[7] Inject Eq. (14)
[8] Summation of telescopic series

Convergence example for smooth function

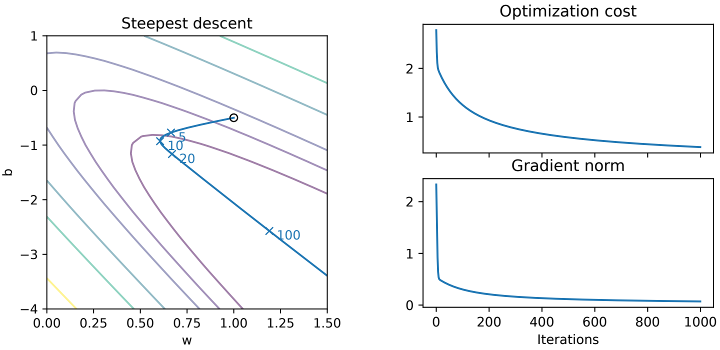




Discussion

- Steepest descent with fixed step $\rho^{(k)} = 0.05$
- Non regularized logistic regression ($\lambda = 0$).
- Slow $O(\frac{1}{k})$ convergence of Gradient Descent.

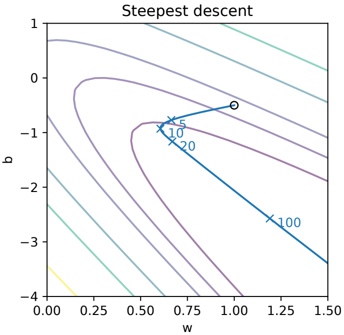
Convergence example for smooth function

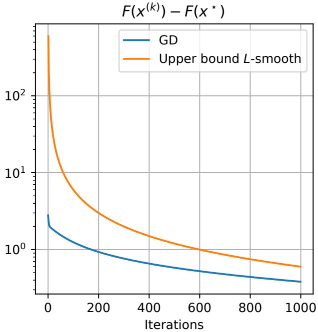


Discussion

- Steepest descent with fixed step $\rho^{(k)} = 0.05$
- Non regularized logistic regression ($\lambda = 0$).
- Slow $O(\frac{1}{k})$ convergence of Gradient Descent.

Convergence example for smooth function

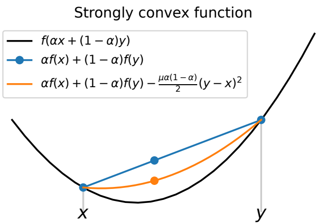


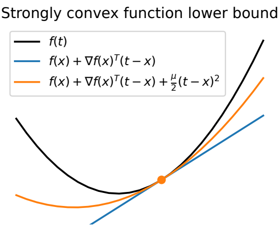


Discussion

- Steepest descent with fixed step $\rho^{(k)} = 0.05$
- Non regularized logistic regression ($\lambda = 0$).
- Slow $O(\frac{1}{k})$ convergence of Gradient Descent.

Assumption 3 : Strong convexity





μ-strongly convex function (recap)

- F is μ-**strongly convex** with $\mu > 0$ if it satisfies $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $0 \leq \alpha \leq 1$

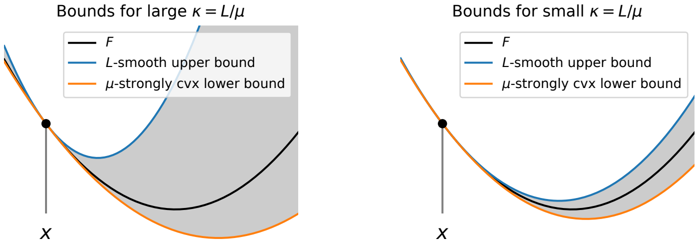
$$F(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha F(\mathbf{x}) + (1-\alpha)F(\mathbf{y}) - \frac{\mu}{2}\alpha(1-\alpha)\|\mathbf{x}-\mathbf{y}\|^2, \qquad (15)$$

- If F is a differentiable μ-strongly convex then

$$F(\mathbf{y}) \geq F(\mathbf{x}) + \nabla F(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|^2, \quad \forall \mathbf{y}, \mathbf{x} \in \text{dom}F$$

- Strongly convex functions have a unique minimum $\mathbf{x}^\star$.

Convergence for strongly convex functions



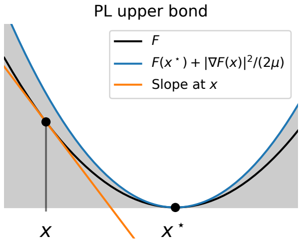
Convergence of gradient descent for μ-strongly convex functions

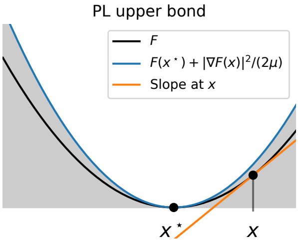
If function F is μ-strongly convex, then the gradient descent with fixed step $\rho^{(k)} = \rho = \frac{1}{L}$ converges to a solution $\mathbf{x}^\star$ of the optimization problem with the following speed:

$$F(\mathbf{x}) - F(\mathbf{x}^*) \leq \left(1 - \frac{\mu}{L}\right)^k \left(F(\mathbf{x}^{(0)}) - F(\mathbf{x}^*)\right) \tag{16}$$

- For a function F, $\mu = \lambda_{\text{min}}(\nabla^2 F(\mathbf{x}))$ and $L = \lambda_{\text{max}}(\nabla^2 F(\mathbf{x}))$.
- The condition $\kappa = \frac{L}{\mu} \geq 1$ has important impact (close to 1 is better approx).
- We say the the gradient descent has a convergence $O(e^{-k/\kappa})$.
- In order to reach a precision ϵ one needs $O(\log(1/\epsilon))$ iterations.

Convergence proof (μ-strongly convex, L-smooth)





Polyak-Lojasciewicz (PL) inequality

If F is a μ-strongly convex function and $\mathbf{x}^\star$ its optimal point then $\forall \mathbf{x}$

$$F(\mathbf{x}) - F(\mathbf{x}^*) \leq \frac{1}{2\mu} \|\nabla F(\mathbf{x})\|^2 \qquad (17)$$

Proof.

Exercise 3 in class. Hints:

- Use strong convexity lower bound.
- Set $\mathbf{y} = \mathbf{x} - \frac{1}{\mu} \nabla F(\mathbf{x})$.
- Inject optimal point $\mathbf{x}^\star$

□

Convergence proof (μ-strongly convex, L-smooth)

$$F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{*}) \leq \left(1 - \frac{\mu}{L}\right)^{k} \|\mathbf{x}^{(0)} - \mathbf{x}^{\star}\|^{2}$$

Proof.

Using the descent lemma (13):

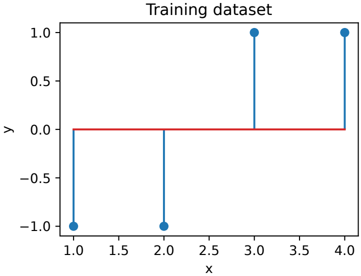
$$\begin{aligned} F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{(k-1)}) &\leq -\frac{1}{2L}\|\nabla F(\mathbf{x}^{(k-1)})\|^{2} \\ &\underset{9}{\leq} -\frac{\mu}{L}\left(F(\mathbf{x}^{(k-1)}) - F(\mathbf{x}^{\star})\right) \end{aligned}$$

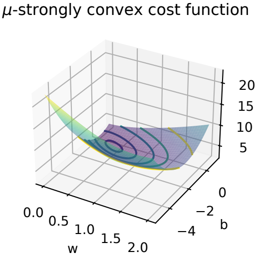
$$\begin{aligned} F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{\star}) &\leq F(\mathbf{x}^{(k-1)}) - F(\mathbf{x}^{\star}) - \frac{\mu}{L}\left(F(\mathbf{x}^{(k-1)}) - F(\mathbf{x}^{\star})\right) \\ &\leq \left(1 - \frac{\mu}{L}\right)\left(F(\mathbf{x}^{(k-1)}) - F(\mathbf{x}^{\star})\right) \\ &\leq \left(1 - \frac{\mu}{L}\right)^{k}\left(F(\mathbf{x}^{(0)}) - F(\mathbf{x}^{\star})\right) \end{aligned}$$

□

[9] Use PL inequality (17)

Convergence example for strongly convex function

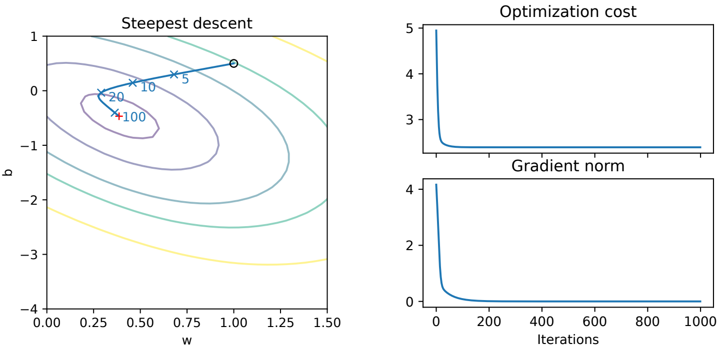




Discussion

- Steepest descent with fixed step $\rho^{(k)} = 0.02$
- Fully regularized logistic regression ($\lambda = 1$ for w and b).
- L-smooth and μ-strongly convex upper bounds.
- Fast $O(e^{-k/\kappa})$ convergence of Gradient Descent.

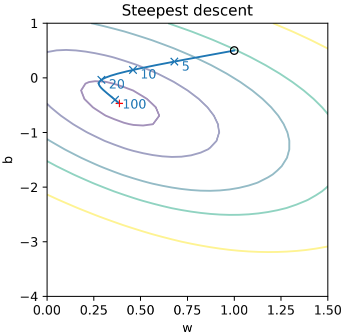
Convergence example for strongly convex function

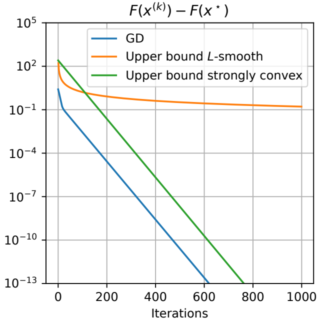


Discussion

- Steepest descent with fixed step $\rho^{(k)} = 0.02$
- Fully regularized logistic regression ($\lambda = 1$ for w and b).
- L-smooth and μ-strongly convex upper bounds.
- Fast $O(e^{-k/\kappa})$ convergence of Gradient Descent.

Convergence example for strongly convex function

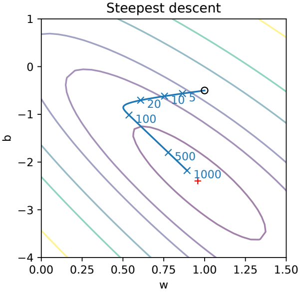


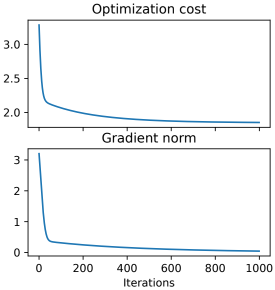


Discussion

- Steepest descent with fixed step $\rho^{(k)} = 0.02$
- Fully regularized logistic regression ($\lambda = 1$ for w and b).
- L-smooth and μ-strongly convex upper bounds.
- Fast $O(e^{-k/\kappa})$ convergence of Gradient Descent.

How to make Gradient Descent faster?





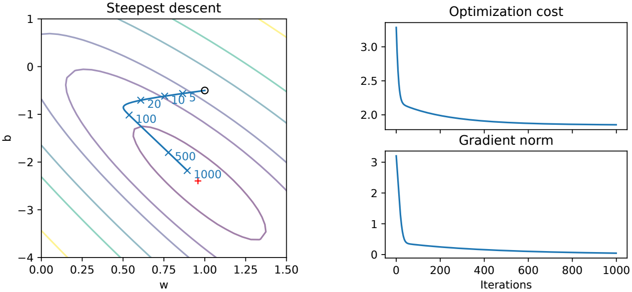
Gradient descent is slow

- Unless on strongly convex fonction it has a $O(\frac{1}{k})$ convergence.
- Needs to recompute the gradient at each iteration ($O(nd)$ in ERM).

Acceleration techniques

- Use adaptive stepsizes (smarter $\rho^{(k)}$).
- Use momentum (remember previous gradients).
- Use second order information (Newton, quasi-Newton).
- Speedup gradient computation (stochastic gradient, slower but more efficient).

How to make Gradient Descent faster?



Gradient descent is slow

- Unless on strongly convex fonction it has a $O(\frac{1}{k})$ convergence.
- Needs to recompute the gradient at each iteration ($O(nd)$ in ERM).

Acceleration techniques

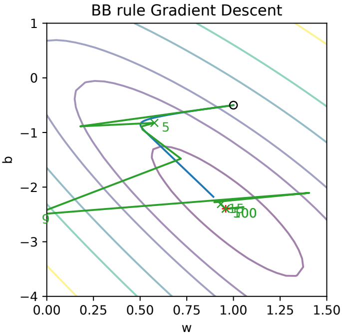
- Use adaptive stepsizes (smarter $\rho^{(k)}$).
- Use momentum (remember previous gradients).
- Use second order information (Newton, quasi-Newton).
- Speedup gradient computation (stochastic gradient, slower but more efficient).

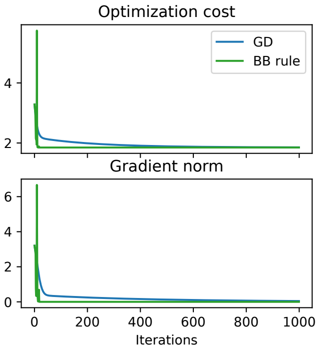
Barzilai-Borwein stepsize (BB-rule)

Principle [Barzilai and Borwein, 1988]

- Use the gradient and the previous gradient to compute the stepsize.
- It is a two-step approximation of the secant method (to cancel the gradient).
- The stepsize is computed as:
  - **Long BB stepsize**:
$$\rho^{(k)} = \frac{\Delta\mathbf{x}^\top \Delta\mathbf{x}}{\Delta\mathbf{x}^\top \Delta\mathbf{g}} \tag{18}$$
  - **Short BB stepsize**:
$$\rho^{(k)} = \frac{\Delta\mathbf{x}^\top \Delta\mathbf{g}}{\Delta\mathbf{g}^\top \Delta\mathbf{g}} \tag{19}$$
- where $\Delta\mathbf{x} = \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}$ and $\Delta\mathbf{g} = \nabla F(\mathbf{x}^{(k)}) - \nabla F(\mathbf{x}^{(k-1)})$.
- The stepsize can be clipped to avoid too large steps (or with linesearch).
- Convergence for quadratic [Raydan, 1993] and non-quadratic functions [Raydan, 1997] with linesearch.
- Variants used for hyperparameter-free optimization with provably better constant.
- Discussed more in details in next courses.

Example of BB rule for Gradient Descent





Discussion

- GD and first step of BB rule use step $\rho^{(k)} = 0.01$.
- Acceleration is important *w.r.t.* steepest descent step.
- Unstable and the stepsize can be too large and lead to loss increase.
- BB rule is best used with linesearch (see next course).

Accelerated gradient descent

Accelerated gradient descent (AGD) [Nesterov, 1983, Walkington, 2023]

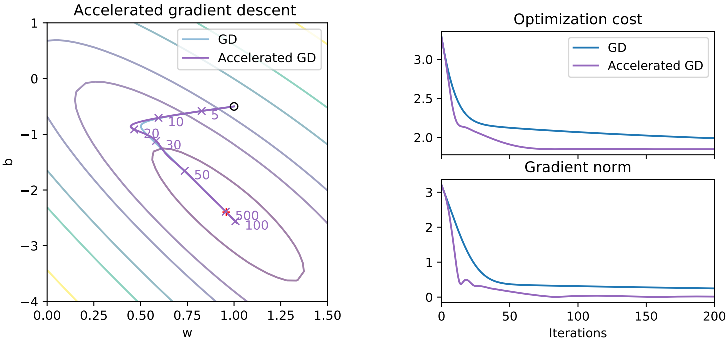
Initialize $\mathbf{x}^{(0)}, \mathbf{y}^{(0)} = \mathbf{x}^{(0)}, \alpha^{(0)} = 0$ and $\rho \leq \frac{1}{L}$
for $k = 0, 1, 2, \ldots$ **do**
    $\mathbf{y}^{(k+1)} \leftarrow \mathbf{x}^{(k)} - \rho \nabla F(\mathbf{x}^{(k)})$
    $\alpha^{(k+1)} =\leftarrow \frac{1+\sqrt{1+4(\alpha^{(k)})^2}}{2}$
    $\mathbf{x}^{(k+1)} \leftarrow \mathbf{y}^{(k+1)} + \frac{\alpha^{(k)}-1}{\alpha^{(k+1)}}(\mathbf{y}^{(k+1)} - \mathbf{y}^{(k)})$
end for

- Also called Nesterov accelerated gradient (NAG).
- Acceleration of gradient descent with momentum.
- Update is gradient step $(\mathbf{y}^{(k+1)})$ + momentum of previous step.
- The algorithm has a $O(\frac{1}{k^2})$ convergence for L-smooth functions and $\rho = \frac{1}{L}$:

$$F(\mathbf{x}^{(k)}) - F(\mathbf{x}^\star) \leq \frac{2L\|\mathbf{x}^{(0)} - \mathbf{x}^\star\|^2}{k^2} \tag{20}$$

- Convergence speed $O(\frac{1}{k^2})$ is optimal for a first order method.

Example of Accelerated Gradient Descent



Discussion

- Both GD and AGD use fixed step $\rho^{(k)} = 0.1$.
- Acceleration speedup is important *w.r.t.* steepest descent step.
- The momentum due the the Nesterov acceleration can be seen in the trajectory.
- Non monotonic convergence but faster than GD.
- Complexity $\mathcal{O}(nd)$ per iteration when no line search.

Least squares and ridge regression

$$\min_{\mathbf{w}} \quad F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^\top \mathbf{x}_i - y_i)^2 + \lambda\|\mathbf{w}\|^2 \tag{21}$$

- Training dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $y_i \in \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}^d$.
- Least Squares ($\lambda = 0$) and Ridge regression ($\lambda > 0$).
- Prediction is done with $\hat{y} = \mathbf{w}^\top \mathbf{x}$.

Exercise 1: Linear regression

1. Reformulate the objective value of least square as a squared norm of residual vector of prediction errors.
2. Compute the gradients for the least square and ridge regression.
3. Express the Hessian and compute the Lipschitz constant L and μ for the least square and ridge regression.

Logistic regression

$$\min_{\mathbf{w}} \quad F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i)) + \lambda \|\mathbf{w}\|^2 \tag{22}$$

- Training dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $y_i \in \{1, 1\}$ and $\mathbf{w} \in \mathbb{R}^d$.
- Regularized logistic regression ($\lambda > 0$).
- Prediction is done with $\hat{y} = \text{sign}(\mathbf{w}^\top \mathbf{x})$.

Exercise 2: Logistic regression

1. Compute the gradients for the logistic regression.
2. Express the Hessian and compute the Lipschitz constant L and μ for the logistic regression.

Lab: Gradient Descent

For the optimization problems

- Least squares regression and Ridge regression.

$$\min_{\mathbf{w}} \quad F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^\top \mathbf{x}_i - y_i)^2 + \lambda\|\mathbf{w}\|^2$$

- Logistic regression.

$$\min_{\mathbf{w}} \quad F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\log(1+\exp(-y_i \mathbf{w}^\top \mathbf{x}_i)) + \lambda\|\mathbf{w}\|^2$$

Your mission

- Implement te loss functions `f` and gradients `df` for the three problems.
- Implement the gradient descent algorithm (and accelerated variant).
- Compare the convergence speed of the three algorithms.

Bibliography I

Convex Optimization [Boyd and Vandenberghe, 2004]

- Available freely online: `https://web.stanford.edu/~boyd/cvxbook/`.

Nonlinear Programming [Bertsekas, 1997]

- Reference optimization book, contains also most of the course.
- Unconstrained optimization (Ch. 1), duality and lagrangian (Ch. 3, 4 ,5).

Convex analysis and monotone operator theory in Hilbert spaces [Bauschke et al., 2011]

- Awesome book with lot's of algorithms, and convergence proofs.
- All definitions (convexity, lower semi continuity) in specific chapters.

Numerical optimization [Nocedal and Wright, 2006]

- Classic introduction to numerical optimization.

References I

Barzilai, J. and Borwein, J. M. (1988).
Two-point step size gradient methods.
IMA Journal of Numerical Analysis, 8(1):141–148.

Bauschke, H. H., Combettes, P. L., et al. (2011).
Convex analysis and monotone operator theory in Hilbert spaces, volume 408.
Springer.

Bertsekas, D. P. (1997).
Nonlinear programming.
Journal of the Operational Research Society, 48(3):334–334.

Boyd, S. and Vandenberghe, L. (2004).
Convex optimization.
Cambridge university press.

Gen, M. and Cheng, R. (1999).
Genetic algorithms and engineering optimization, volume 7.
John Wiley & Sons.

References II

Hunter, D. R. and Lange, K. (2004).
A tutorial on mm algorithms.
The American Statistician, 58(1):30–37.

Kennedy, J. and Eberhart, R. (1995).
Particle swarm optimization.
In *Proceedings of ICNN'95-international conference on neural networks*, volume 4, pages 1942–1948. ieee.

Nelder, J. A. and Mead, R. (1965).
A simplex method for function minimization.
The computer journal, 7(4):308–313.

Nesterov, Y. E. (1983).
A method for solving the convex programming problem with convergence rate o (1/kˆ 2).
In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547.

Nocedal, J. and Wright, S. (2006).
Numerical optimization.
Springer Science & Business Media.

References III

Raydan, M. (1993).
On the barzilai and borwein choice of steplength for the gradient method.
IMA Journal of Numerical Analysis, 13(3):321–326.

Raydan, M. (1997).
The barzilai and borwein gradient method for the large scale unconstrained minimization problem.
SIAM Journal on Optimization, 7(1):26–33.

Sun, Y., Babu, P., and Palomar, D. P. (2016).
Majorization-minimization algorithms in signal processing, communications, and machine learning.
IEEE Transactions on Signal Processing, 65(3):794–816.

Walkington, N. J. (2023).
Nesterov's method for convex optimization.
SIAM Review, 65(2):539–562.

Wolpert, D. H. and Macready, W. G. (1997).
No free lunch theorems for optimization.
IEEE transactions on evolutionary computation, 1(1):67–82.