Optimization for Machine Learning

Part II : Introduction to SGD

Lionel Tondji

African Master's in Machine Intelligence

July 24, 2024

Structure of Optimization Problems Arising in Training Supervised machine Learning Models

Optimization Problems Arising in Machine Learning

In this course, we are interested in the optimization problem

$$\min_{w \in \mathbb{R}^d} f(w) + R(w) \tag{1}$$

Optimization Problems Arising in Machine Learning

In this course, we are interested in the optimization problem

$$\min_{w \in \mathbb{R}^d} f(w) + R(w) \tag{1}$$

Typical structure of f:

- Infinite sum

$$f(w) = E_{\zeta \sim \mathcal{D}}[f_\zeta(w)] \tag{2}$$

Optimization Problems Arising in Machine Learning

In this course, we are interested in the optimization problem

$$\min_{w \in \mathbb{R}^d} f(w) + R(w) \tag{1}$$

Typical structure of f:

- Infinite sum

$$f(w) = E_{\zeta \sim \mathcal{D}}[f_\zeta(w)] \tag{2}$$

- Finite sum :

$$f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w) \tag{3}$$

Optimization Problems Arising in Machine Learning

In this course, we are interested in the optimization problem

$$\min_{w \in \mathbb{R}^d} f(w) + R(w) \tag{1}$$

Typical structure of f:

- Infinite sum

$$f(w) = E_{\zeta \sim \mathcal{D}}[f_\zeta(w)] \tag{2}$$

- Finite sum :

$$f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w) \tag{3}$$

- Finite sum of Finite Sums :

$$f_i(w) = \frac{1}{m} \sum_{j=1}^{m} f_{ij}(w) \tag{4}$$

These problems are of keys importance in supervised learning theory and pratice.

Common feature: It is prohibitively expensive to compute the gradient of f, while an unbiased estimator of the gradient can be computed efficiently/cheaply.

Stochastic Optimization and Machine Learning

In the stochastic optimization problem

$$f(w) = E_{\zeta \sim \mathcal{D}}[f_\zeta(w)]$$

- w represents a machine learning model described by d parameters/features (e.g logistic regression or a deep neural network).

Stochastic Optimization and Machine Learning

In the stochastic optimization problem

$$f(w) = E_{\zeta \sim \mathcal{D}}[f_\zeta(w)]$$

- w represents a machine learning model described by d parameters/features (e.g logistic regression or a deep neural network).
- $\mathcal{D}$ is an unknown distribution of labelled examples,

Stochastic Optimization and Machine Learning

In the stochastic optimization problem

$$f(w) = E_{\zeta \sim \mathcal{D}}[f_\zeta(w)]$$

- w represents a machine learning model described by d parameters/features (e.g logistic regression or a deep neural network).
- $\mathcal{D}$ is an unknown distribution of labelled examples,
- $f_\zeta(w)$ represents the loss of model w on a data point ζ, and

Stochastic Optimization and Machine Learning

In the stochastic optimization problem

$$f(w) = E_{\zeta \sim \mathcal{D}}[f_\zeta(w)]$$

- w represents a machine learning model described by d parameters/features (e.g logistic regression or a deep neural network).
- $\mathcal{D}$ is an unknown distribution of labelled examples,
- $f_\zeta(w)$ represents the loss of model w on a data point ζ, and
- f is the generalization error.

Problem (1) seeks to find the model w minimizing the generalization error

1. In statistical learning theory one assumes that while $\mathcal{D}$ is not known, samples $\zeta \sim \mathcal{D}$ are available.
2. In such case, $\nabla f(w)$ is not computable, while $\nabla f_\zeta(w)$, which is an unbiased estimator of the gradient of f at w, is easily computable.

Finite Sum Problems

In this course we will focus on functions f which arise as averages of very large number of (smooth) functions:

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$$

Finite Sum Problems

In this course we will focus on functions f which arise as averages of very large number of (smooth) functions:

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$$

- This problem often arises by approximation of the stochastic optimization loss function (2) via Monte Carlo Integration.
- Known as the empirical risk minimization (ERM) problem.
- ERM is currently the dominant paradigm for solving supervised learning problems.

Finite Sum Problems

In this course we will focus on functions f which arise as averages of very large number of (smooth) functions:

$$f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

- This problem often arises by approximation of the stochastic optimization loss function (2) via Monte Carlo Integration.
- Known as the empirical risk minimization (ERM) problem.
- ERM is currently the dominant paradigm for solving supervised learning problems.
- If index i is chosen uniformly at random from $[n] = \{1, 2, \dots, n\}$, $\nabla f_i(w)$ is an unbiased estimator of $\nabla f(w)$.
- Typically, $\nabla f_i(w)$ is about n times less expensive to compute than $\nabla f(w)$.

Distributed Training

In distributed of supervised models, one considers the finite sum problem (3), with n being the number of machines, and each f_i

- also having a finite sum structure, i.e,

$$f_i(w) = \frac{1}{m} \sum_{j=1}^{m} f_{ij}(w) \tag{5}$$

where m corresponds to the number of training examples stored on machine i.

Distributed Training

In distributed of supervised models, one considers the finite sum problem (3), with n being the number of machines, and each f_i

- also having a finite sum structure, i.e,

$$f_i(w) = \frac{1}{m}\sum_{j=1}^{m} f_{ij}(w) \tag{5}$$

  where m corresponds to the number of training examples stored on machine i.

- or an infinite-sum structure, i.e,

$$f_i(w) = E_{\zeta_i \sim \mathcal{D}_i}[f_{i\zeta_i}(w)] \tag{6}$$

  where $\mathcal{D}_i$ is the distribution of data stored on machine i.

SGD

Stochastic gradient descent (SGD) is a state-of-the-art algorithmic paradigm for solving optimization problem (1) in situations where f is either of structure (2) or (3).
In its generic form (proximal) SGD defines the new iterate by subtracting a multiple of a stochastic gradient from the current iterate, and subsequently applying the proximal operator of R:

$$w_{k+1} = prox_{\gamma R}(w_k - \gamma g^k) \tag{7}$$

SGD

Stochastic gradient descent (SGD) is a state-of-the-art algorithmic paradigm for solving optimization problem (1) in situations where f is either of structure (2) or (3).
In its generic form (proximal) SGD defines the new iterate by subtracting a multiple of a stochastic gradient from the current iterate, and subsequently applying the proximal operator of R:

$$w_{k+1} = prox_{\gamma R}(w_k - \gamma g^k) \tag{7}$$

- g^k is an unbiased estimator of the gradient (i.e a "stochastic gradient"):

$$E[g^k / w_k] = \nabla f(w_k) \tag{8}$$

SGD

Stochastic gradient descent (SGD) is a state-of-the-art algorithmic paradigm for solving optimization problem (1) in situations where f is either of structure (2) or (3).
In its generic form (proximal) SGD defines the new iterate by subtracting a multiple of a stochastic gradient from the current iterate, and subsequently applying the proximal operator of R:

$$w_{k+1} = prox_{\gamma R}(w_k - \gamma g^k) \tag{7}$$

- g^k is an unbiased estimator of the gradient (i.e a "stochastic gradient"):
$$E[g^k / w_k] = \nabla f(w_k) \tag{8}$$

- $$prox_R(x) = arg \min_u \left\{ R(u) + \frac{1}{2}\|u - x\|^2 \right\}$$

The Prox Operator

Some facts about the prox operator[1]:

1. single-valuedness: $x \mapsto prox_R(x)$ is a function
2. non-expansiveness:

$$\|prox_R(x) - prox_R(y)\| \leq \|x - y\|, \quad \forall x, y \in \mathbb{R}^d$$

3. Moreau decomposition:

$$prox_R(x) - prox_{R^*}(x) = x, \quad \forall x \in \mathbb{R}^d$$

Here R^* is the Fenchel conjugate[2] of R.

[1] Assume $R : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is proper, closed and convex.
[2] $R^*(x) = sup_{y \in \mathbb{R}^d} \{\langle x, y \rangle - R(y)\}$

Stochastic Gradient

There are infinitely many ways of obtaining a random vector g^k satisfying (8)

- Prox: flexibility to construct stochastic gradients in various ways based on problem structure , and in order to target desirable properties such as:
  - convergence speed
  - iteration cost
  - overall complexity
  - parallelizability
  - suitability for given computing architecture
  - communication cost
  - generalization properties

Stochastic Gradient

There are infinitely many ways of obtaining a random vector g^k satisfying (8)

- Cons: A crazy ZOO of methods
  - Little hard to get into the fields, hard to keep up with new results
  - Considerable challenges in terms of convergence analysis. Indeed, if one aims to, as one should, obtain the sharpest bounds possible, dedicated analyses are needed to handle each of the particular variants of SGD.

Batch SGD = Gradient Descent

Gradient Descent

We first describe the (proximal) gradient descent (GD) method for solving the regularized convex optimization problem

$$\min_{w \in \mathbb{R}^d} f(w) + R(w) \tag{9}$$

This is the most basic of all SGD methods, and a starting point for the development of more elaborate variants.

Algorithm GD

starting points $x_0 \in \mathbb{R}^d$, **learning rate** $\gamma > 0$
for $k = 0, 1, 2, \cdots$ **do**
  Set $g^k = \nabla f(w_k)$
  $w_{k+1} = prox_{\gamma R}(w_k - \gamma g^k)$
end for

The idea is f might be something complicated but the linear approximation is simple.

$$\begin{aligned} f(w) &\simeq \text{linear function} \\ &= f(w_0) + \langle \nabla f(w_0), w - w_0 \rangle \end{aligned}$$

The idea is f might be something complicated but the linear approximation is simple.

$$\begin{aligned} f(w) &\simeq \text{linear function} \\ &= f(w_0) + \langle \nabla f(w_0), w - w_0 \rangle \end{aligned}$$

GD is an iterative algorithm where at each step we minimize a linear approximation but plus an additional term that makes sure we do not roll all the way down hill.

The idea is f might be something complicated but the linear approximation is simple.

$$\begin{aligned} f(w) &\simeq \text{linear function} \\ &= f(w_0) + \langle \nabla f(w_0), w - w_0 \rangle \end{aligned}$$

GD is an iterative algorithm where at each step we minimize a linear approximation but plus an additional term that makes sure we do not roll all the way down hill.

Idea of GD:

- start w_0
- $w_{k+1} = arg \min_w f(w_k) + \langle \nabla f(w_k), w - w_k \rangle + \frac{1}{2\gamma} \|w - w_k\|_2^2$

* 1^{st} term: linear function

* 2^{nd} term : quadratic term that penalize w being very far from w_k.

Taking deriv (grad) set to zero:

$$0 + \nabla f(w_k) + \frac{1}{\gamma}(w - w_k) = 0$$

$$\Rightarrow w_{k+1} = w_k - \gamma \nabla f(w_k)$$

Example: $f(x) = 3x^2 + 4x - 2$

Taking deriv (grad) set to zero:

$$0 + \nabla f(w_k) + \frac{1}{\gamma}(w - w_k) = 0$$

$$\Rightarrow w_{k+1} = w_k - \gamma \nabla f(w_k)$$

Example: $f(x) = 3x^2 + 4x - 2$

1. Can solve this directly: $6x + 4 = 0 \Rightarrow x^* = -\frac{2}{3}$

Taking deriv (grad) set to zero:

$$0 + \nabla f(w_k) + \frac{1}{\gamma}(w - w_k) = 0$$

$$\Rightarrow w_{k+1} = w_k - \gamma \nabla f(w_k)$$

Example: $f(x) = 3x^2 + 4x - 2$

1. Can solve this directly: $6x + 4 = 0 \Rightarrow x^* = -\frac{2}{3}$
2. Apply gradient descent. Initialize at x_1, take step size γ.
   - GD iteration: $x^+ = x - \gamma \nabla f(x) = x - \gamma(6x + 4)$

Taking deriv (grad) set to zero:

$$0 + \nabla f(w_k) + \frac{1}{\gamma}(w - w_k) = 0$$

$$\Rightarrow w_{k+1} = w_k - \gamma \nabla f(w_k)$$

Example: $f(x) = 3x^2 + 4x - 2$

1. Can solve this directly: $6x + 4 = 0 \Rightarrow x^* = -\frac{2}{3}$
2. Apply gradient descent. Initialize at x_1, take step size γ.
   - GD iteration: $x^+ = x - \gamma\nabla f(x) = x - \gamma(6x + 4)$

$$\begin{aligned}
\Rightarrow x_{k+1} &= x_k - \gamma 6 x_k - 4\gamma \\
&= (1 - 6\gamma)x_k - 4\gamma \\
&\vdots \\
&= (1-6\gamma)^k x_1 - \big((1-6\gamma)^{k-1} + (1-6\gamma)^{k-2} + \cdots + 1\big)4\gamma \\
&= (1-6\gamma)^k x_1 - \frac{1-(1-6\gamma)^k}{1-(1-6\gamma)}4\gamma
\end{aligned}$$

Need $|1-6\gamma| < 1 \Rightarrow (1-6\gamma)^k \to 0 \quad \text{as } k \to \infty.$

$$\begin{aligned} x_{k+1} &= (1-6\gamma)^k x_1 + \frac{(1-6\gamma)^k}{6\gamma} - \frac{1}{6\gamma} 4\gamma \\ &= (1-6\gamma)^k \left[x_1 + \frac{2}{3} \right] - \frac{2}{3} \to -\frac{2}{3} \text{as } k \text{ grows} \end{aligned}$$

$x_k \to -\frac{2}{3}$ very quickly with linear rate.

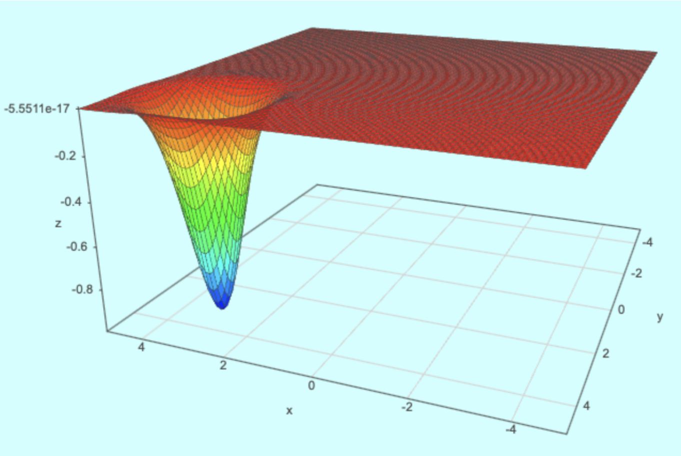
Need $|1-6\gamma|<1 \Rightarrow (1-6\gamma)^k \to 0 \quad \text{as } k \to \infty.$

$$
\begin{aligned}
x_{k+1} &= (1-6\gamma)^k x_1 + \frac{(1-6\gamma)^k}{6\gamma} - \frac{1}{6\gamma}4\gamma \\
&= (1-6\gamma)^k\left[x_1+\frac{2}{3}\right] - \frac{2}{3} \to -\frac{2}{3}\text{as } k \text{ grows}
\end{aligned}
$$

$x_k \to -\frac{2}{3}$ very quickly with linear rate.

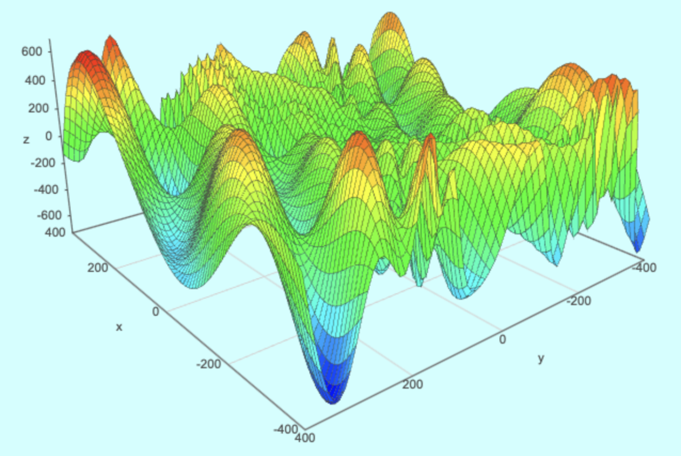
1. **Good News**: GD for this function, converges very quickly. Error goes down with $(1-6\gamma)^k$ fast when $\gamma < \frac{1}{6}$.
2. **Step size**: small enough so that $|1-6\gamma|<1$
3. **Improvement at every iteration**: Exercise: check that $f(w_{k+1}) \leq f(w_k)$

Optimization is hard (in general)



$$f(x, y) = -cos(x)cos(y)exp(-(x-\pi)^2 - (y-\pi)^2) \quad \text{in } [-5, 5]^2$$

Optimization is hard (in general)

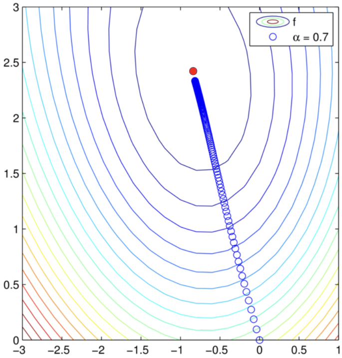


$f(x,y) =$
$-(y+47)sin\sqrt{|\frac{x}{2}+(y+47)|} - xsin\sqrt{|\frac{x}{2}-(y+47)|}$ in $[-400,400]^2$

Gradient Descent Example

A Logistic Regression problem using the fourclass labelled data from LIBSVM $(n, d) = (862, 2)$.
Logistic Regression :

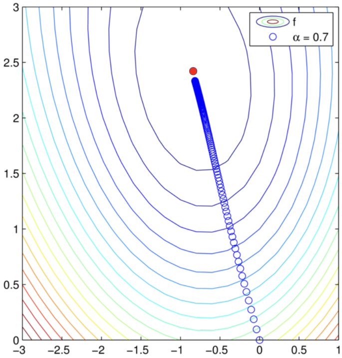
$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} ln(1 + e^{-y^i \langle w, x^i \rangle}) + \lambda \|w\|_2^2$$



Gradient Descent Example

A Logistic Regression problem using the fourclass labelled data from LIBSVM $(n, d) = (862, 2)$.
Logistic Regression :

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} ln(1 + e^{-y^i \langle w, x^i \rangle}) + \lambda \|w\|_2^2$$

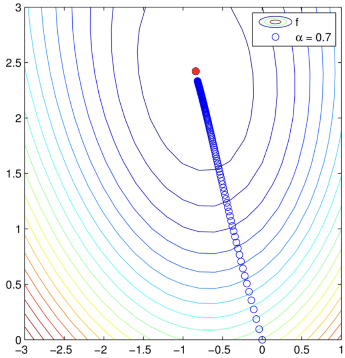


1. Can we prove that this always works?

Gradient Descent Example

A Logistic Regression problem using the fourclass labelled data from LIBSVM $(n, d) = (862, 2)$.
Logistic Regression :

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} ln(1 + e^{-y^i \langle w, x^i \rangle}) + \lambda \|w\|_2^2$$



1. Can we prove that this always works?
2. **No**! There is no universal optimization method. The "no free lunch" of Optimization
3. Need assumptions: **Convex** and **smooth** training problems

Main assumption

Nice property:

$$\text{If } \nabla f(w^*) = 0 \quad \text{then } f(w^*) \leq f(w), \ \forall \ w \in \mathbb{R}^d$$

$\Rightarrow$ All stationary points are global minima.

Main assumption

Nice property:

$$\text{If } \nabla f(w^*) = 0 \quad \text{then } f(w^*) \leq f(w), \ \forall \ w \in \mathbb{R}^d$$

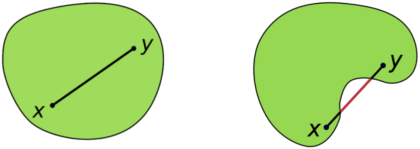
$\Rightarrow$ All stationary points are global minima.

Lemma

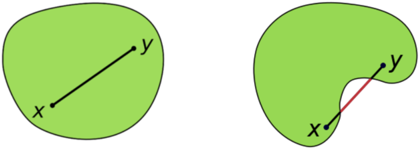
Convexity $\Rightarrow$ ***Nice property****.*
If $f(w) \geq f(y) + \langle \nabla f(y), w - y \rangle, \quad \forall \ w, y \in \mathbb{R}^d$ *then Nice property holds.*

PROOF: Choose $y = w^*$.

Convex sets - Definition



Convex sets - Definition



A set $\mathbb{C} \subseteq \mathbb{R}^n$, is convex if

$$\forall\, x, y \in \mathbb{C},\ \forall\ \lambda \in [0,1], \quad \lambda x + (1-\lambda) y \in \mathbb{C}. \tag{10}$$

Why it is important ?

Definition

$M \in \mathbb{R}^{n \times n}$ matrix is p.s.d if:

1. Symmetric
2. $x^\top M x \geq 0, \ \forall \ x \in \mathbb{R}^n$

We denote by $\mathcal{S}_+^n$, the set of symmetric p.s.d matrices.

The set of p.s.d matrices is a convex set.
Let $M_1, M_2 \in \mathcal{S}_+^n$, we want to show that
$M = \lambda M_1 + (1 - \lambda) M_2 \in \mathcal{S}_+^n, \ \lambda \in [0, 1]$

A symmetric matrix M is **copositive** if

$$x^\top Mx \geq 0,\ \forall\ x \in \mathbb{R}_+^n$$

We denote by $\mathcal{C}^n$, the set of symmetric copositive matrices.

1. Q: Is the set of copositives matrices bigger or smaller than $\mathcal{S}_+^n$?
2. In general it is intractable to determine whether a matrix M is copositive.

Example

1. Every matrix with only non negative entries is copositive Hence $M = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is copositive but not p.s.d due to $det(M) = -1$.
2. Every p.s.d is also copositive but the converse is false i.e. $\mathcal{S}_+^n \subseteq \mathcal{C}^n$

Exercise

Characterize the triples $(x, y, z) \in \mathbb{R}^3$ for which the matrix $M = \begin{pmatrix} x & z \\ z & y \end{pmatrix}$ is

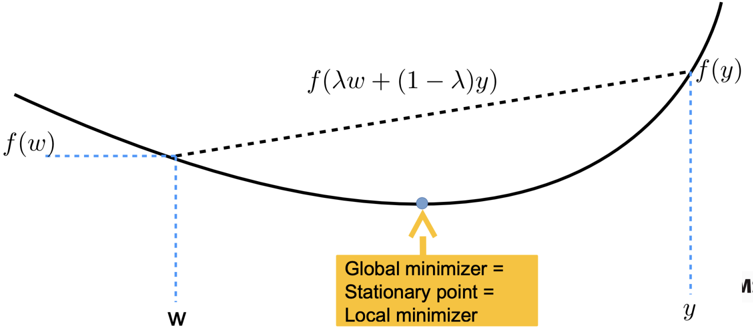
1. copositive
2. p.s.d

Definition

We say that $f : dom(f) \subseteq \mathbb{R}^d \to \mathbb{R}$ is **convex** if $dom(f)$ is **convex** and

$$f(\lambda w + (1-\lambda)y) \leq \lambda f(w) + (1-\lambda)f(y), \tag{11}$$

$\forall\ w, y \in dom(f),\ \lambda \in [0,1]$

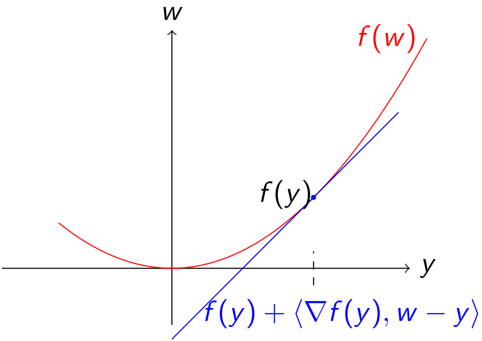


Def 2 : Convexity - First derivative

Definition

A differential function $f : dom(f) \subseteq \mathbb{R}^d \to \mathbb{R}$ is **convex** iff

$$f(w) \geq f(y) + \langle \nabla f(y), w - y \rangle, \quad \forall\, w, y \in dom(f) \tag{12}$$

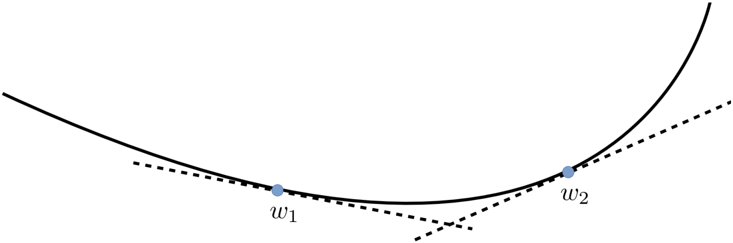


Def 3 : Convexity - Second derivative

Definition

A twice differentiable function $f : dom(f) \subseteq \mathbb{R}^d \to \mathbb{R}$ is **convex** iff

$$\nabla^2 f(w) \succcurlyeq 0 \Leftrightarrow v^\top \nabla^2 f(w) v \geq 0, \quad \forall\ w, v \in dom(f) \tag{13}$$



$w_1 \leq w_2 \Rightarrow f^{'}(w_1) \leq f^{'}(w_2)$.
($\nabla^2 f(w)$ p.s.d i.e. all eigen values of $\nabla^2 f(w)$ are ≥ 0).

Convexity : Examples

1. Norms and squared norms : $x \mapsto \|x\|$, $x \mapsto \|x\|^2$
2. Negative log and logistic : $x \mapsto \log(x)$, $x \mapsto log(1 + e^{-y\langle a,x\rangle})$
3. Hinge Loss : $x \mapsto max\{0, 1 - yx\}$
4. Negatives log determinant, exponentiation ... etc

Def 2' : Convex functions (non-smooth)

Definition

A function f is **convex** if $\forall\, y,\ \exists g$ such that

$$f(w) \geq f(y) + \langle g, w - y \rangle \tag{14}$$

i.e at every point of the function there exists some linear function which touch the function at that point but it is not necessarily unique as a gradient.

- If f is convex and differentiable, then $g = \nabla f(y)$ satisfies this. It is unique.

Subgradient and Subdifferential

Definition

For a convex function f, a vector g such that

$$f(w) \geq f(y) + \langle g, w - y \rangle, \quad \forall y$$

is called a subgradient.

The set of subgradients of f at a point w is called the subdifferential of f at w and denoted by $\partial f(w)$

More examples

1. Applying the three definitions for $f(x) = x^\top Qx$, $Q \succcurlyeq 0$
2. The max of convex function is convex.
3. The min may not be.
4. Largest element of a vector :

$$f(x_1, \dots, x_n) = \text{maximum element (defined on } \mathbb{R}^n_+)$$

More examples

1. Applying the three definitions for $f(x) = x^\top Qx$, $Q \succcurlyeq 0$
2. The max of convex function is convex.
3. The min may not be.
4. Largest element of a vector :

$$f(x_1, \ldots, x_n) = \text{maximum element (defined on } \mathbb{R}^n_+)$$

In fact

$$\begin{aligned} f(x_1, \ldots, x_n) &= f(x) \\ &= \max\{e_1^\top x, e_2^\top x, \ldots, e_n^\top x\} \\ &= \max_{1 \le i \le n} f_i(x) \end{aligned}$$

each $f_i(x)$ is convex.

More example : The largest eigenvalue of a symmetric matrix

The function f defined by:

$$f(Q) = \lambda_{\max}(Q),$$

is convex. We can show that

$$\lambda_{\max}(Q) = \text{max of convex functions}$$

Recall: Q symmetric $\Rightarrow x^\top Qx \leq \lambda_{\max}\|x\|_2^2$ and $x^\top Qx = \lambda_{\max}\|x\|_2^2$ when

$$x = \text{eigen vector corresponding to } \lambda_{\max}$$

because $x^\top Qx = x^\top(\lambda_{\max}x) = \lambda_{\max}x^\top x = \lambda_{\max}\|x\|_2^2$

$$\lambda_{\max} = \sup_{\|x\|_2=1} x^\top Qx$$

$x^\top Qx$ as a function of Q is linear since $\langle xx^\top, Q\rangle = x^\top Qx$

Monotonicity

If f is convex, the gradient (subgradient) of f is monotone i.e

- $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$
- $\langle g_x - g_y, x - y \rangle \geq 0$ for $g_x \in \partial f(x), g_y \in \partial f(y)$

Proof:

Monotonicity

If f is convex, the gradient (subgradient) of f is monotone i.e

- $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$
- $\langle g_x - g_y, x - y \rangle \geq 0$ for $g_x \in \partial f(x), g_y \in \partial f(y)$

Proof:

$$f(y) \geq f(x) + \langle g_x, y - x \rangle$$
$$f(x) \geq f(y) + \langle g_y, x - y \rangle$$
$$f(x) + f(y) \geq f(x) + f(y) + \langle g_x - g_y, y - x \rangle$$

$\Rightarrow \langle g_x - g_y, y - x \rangle \leq 0 \Rightarrow \langle g_x - g_y, x - y \rangle \geq 0$

Equivalence : Convexity and monotonicity

a) If f is convex (def 3), its gradient is monotone.
b) If the gradient of f is monotone, then f is convex (def 2)

Proof.

a) Recall that $\int_0^1 F^{'}(t)dt = F(1) - F(0)$ and by assumption $\nabla^2 f(x) \succcurlyeq 0$

$$\begin{aligned}\int_0^1 (x-y)^\top \nabla^2 f(tx + (1-t)y)dt &= \int_0^1 \frac{d}{dt}(\nabla f(t(x-y)+y)dt) \\ &= \nabla f(x) - \nabla f(y)\end{aligned}$$

Taking the inner product with $(x - y)$

$$\int_0^1 (x-y)^\top \nabla^2 f(tx + (1-t)y)(x-y)dt = \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$$

Equivalence : Convexity and monotonicity

$$b)\quad \int_0^1 \nabla f((y-x)t+x)^\top (y-x)dt = \int_0^1 \frac{d}{dt} f((y-x)t+x)dt$$
$$= f(y) - f(x)$$

$$\Rightarrow f(y) = f(x) + \int_0^1 \nabla f((y-x)t+x)^\top (y-x)dt \ \left(\geq \langle \nabla f(x), y-x \rangle \right)$$

We can show this inequality holds, by showing that the integral is smallest at $t=0$. We will use monotone property.

$$\langle \nabla f((y-x)t+x) - \nabla f(x), (y-x)t + x - x \rangle \geq 0$$
$$\langle \nabla f((y-x)t+x) - \nabla f(x), (y-x) \rangle \geq 0$$

Let $h(t) = \nabla f((y-x)t+x)^\top (y-x)$, then
$h(t) - h(0) = \langle \nabla f((y-x)t+x) - \nabla f(x), (y-x) \rangle \geq 0$

Equivalence : Convexity and monotonicity

$$h(t) = \nabla f((y-x)t + x)^\top (y-x),$$

then

$$h(t) - h(0) = \langle \nabla f((y-x)t + x) - \nabla f(x), (y-x) \rangle \geq 0$$

$$\Rightarrow \int h(t)dt \geq h(0) \cdot 1$$

$$\begin{aligned} \Rightarrow f(y) &= f(x) + \int_0^1 h(t)dt \\ &\geq f(x) + h(0) \\ &= f(x) + \langle \nabla f(x), y - x \rangle \end{aligned}$$

Example

Sum of squares: Let $a_1, a_2, \ldots, a_n$, find x to minimize : $\frac{1}{n}\sum_{i=1}^{n}(a_i - x)^2$.

Example

Sum of squares: Let $a_1, a_2, \ldots, a_n$, find x to minimize : $\frac{1}{n}\sum_{i=1}^{n}(a_i - x)^2$.
Take derivative, set to 0

$$-\frac{2}{n}\sum_{i=1}^{n}(a_i - x) = 0 \Rightarrow \sum_{i=1}^{n}(a_i - x) = 0$$
$$\Rightarrow \hat{x} = \frac{1}{n}\sum_{i=1}^{n} a_i$$

Optimality conditions for convex Optimization

Example

Ridge regression: Let $(y_i, x_i)_{1\le i\le n}, x_i \in \mathbb{R}^d$, find β to minimize :

$$f(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2 + \lambda \sum_{i=1}^{n} \beta_i^2$$

.

Optimality conditions for convex Optimization

Example

Ridge regression: Let $(y_i, x_i)_{1 \leq i \leq n}, x_i \in \mathbb{R}^d$, find β to minimize :

$$f(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^\top \beta)^2 + \lambda \sum_{i=1}^{n} \beta_i^2$$

.

$f(\beta) = \frac{1}{n}\|X\beta - y\|_2^2 + \lambda\|\beta\|_2^2$, take derivative, set to 0

$$\begin{aligned}\nabla f(\beta) = 0 &\Rightarrow \frac{2}{n}X^\top(X\beta - y) + 2\lambda\beta = 0 \\ &\Rightarrow (\frac{1}{n}X^\top X + \lambda I)\beta = \frac{1}{n}X^\top y \\ &\Rightarrow \beta = (X^\top X + \lambda n I)^{-1}X^\top y\end{aligned}$$

Non-differentiable functions

Let f be a convex function and we want to solve $\min_x f(x)$.

Definition

$\hat{x}$ is **an** optimal solution if $0 \in \partial f(\hat{x})$

$$g_x \in \partial f(x) \Leftrightarrow f(y) \geq f(x) + \langle g_x, y - x \rangle, \quad \forall y$$

If $0 \in \partial f(\hat{x})$, the definition reads:

$$f(y) \geq f(\hat{x}) + 0, \quad \forall y$$

i.e. $\hat{x}$ is **an** optimal solution.

Example : **Sum of absolute values**

Example

Let $a_1, a_2, \ldots, a_n$, find x to minimize : $\frac{1}{n}\sum_{i=1}^{n} |a_i - x|$.
$\hat{x}$ optimal if $0 \in \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \hat{x}}(|a_i - x|)$

Note that

$$\frac{\partial}{\partial z}|z| = \begin{cases} -1, & \text{if z<0}, \\ 1, & \text{if z>0}, \\ [-1, 1], & \text{if z=0} \end{cases}$$

Exercise: Apply this definition of subdifferential to the above problem to find the solution.

Example

Let $a_1, a_2, \ldots, a_n$, find x to minimize : $\frac{1}{n}\sum_{i=1}^{n}|a_i - x|$.
$\hat{x}$ optimal if $0 \in \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \hat{x}}(|a_i - x|)$

Note that

$$\frac{\partial}{\partial z}|z| = \begin{cases} -1, & \text{if z<0,} \\ 1, & \text{if z>0,} \\ [-1, 1], & \text{if z=0} \end{cases}$$

Exercise: Apply this definition of subdifferential to the above problem to find the solution.
Answer: $\hat{x} = \text{median}\{a_1, a_2, \ldots, a_n\}$

Example : **Lasso regression**

Example

Data: Let $\{(y_i, x_i)_{1\le i\le n}, y_i \in \mathbb{R}, x_i \in \mathbb{R}^d\}$.
Find β to minimize :

$$f(\beta) = \frac{1}{n}\|X\beta - y\|_2^2 + \lambda\|\beta\|_1, \quad \|\beta\|_1 = \sum_{i=1}^{n} |\beta_i|.$$

$$\beta \text{ optimal if } 0 \in \frac{\partial}{\partial\beta}(\frac{1}{n}\|X\beta - y\|_2^2 + \lambda\|\beta\|_1)$$

$$0 \in \frac{2}{n}X^\top(X\beta - y) + \lambda\frac{\partial}{\partial\beta}(\|\beta\|_1) = 0$$

Example : **Lasso regression**

Example

Data: Let $\{(y_i, x_i)_{1\leq i\leq n}, y_i \in \mathbb{R}, x_i \in \mathbb{R}^d\}$.
Find β to minimize :

$$f(\beta) = \frac{1}{n}\|X\beta - y\|_2^2 + \lambda\|\beta\|_1, \quad \|\beta\|_1 = \sum_{i=1}^{n} |\beta_i|.$$

$$\beta \text{ optimal if } 0 \in \frac{\partial}{\partial\beta}(\frac{1}{n}\|X\beta - y\|_2^2 + \lambda\|\beta\|_1)$$
$$0 \in \frac{2}{n}X^\top(X\beta - y) + \lambda\frac{\partial}{\partial\beta}(\|\beta\|_1) = 0$$

$\hat{\beta}$ is optimal if $\exists z$ with $z_i = \begin{cases} sign(\hat{\beta}_i), & \text{if } \hat{\beta}_i \neq 0, \\ [-1, 1], & \text{if } \hat{\beta}_i = 0 \end{cases}$ and
$0 = \frac{2}{n}X^\top(X\beta - y) + \lambda z.$

Smoothness : "Self tuning" property

Definition

We say that $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is **smooth** with constant L if

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \cdot \|x - y\|_2, \quad \forall\ x, y \in \mathbb{R}^d \tag{15}$$

If a twice differentiable $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is L-smooth then

1. $v^\top \nabla^2 f(x) v \leq L \cdot \|v\|_2^2, \quad \forall\ x, v \in \mathbb{R}^d.$
2. $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2, \quad \forall\ x, y \in \mathbb{R}^d$

"Self tuning property": $\|\nabla f(x)\| \to 0$, as $x \to x^*$.

Smoothness

Fact: f is L-smooth iff $g(x) = \frac{L}{2}\|x\|_2^2 - f(x)$ is convex.
recall : monotone $\Leftrightarrow$ convex, so it is enough to show that ∇g is monotone.

Proof.

$\Rightarrow$) we want to proof that $\langle \nabla g(x) - \nabla g(y), x - y\rangle \geq 0$.

Smoothness

Fact: f is L-smooth iff $g(x) = \frac{L}{2}\|x\|_2^2 - f(x)$ is convex.
recall : monotone $\Leftrightarrow$ convex, so it is enough to show that ∇g is monotone.

Proof.

$\Rightarrow$) we want to proof that $\langle \nabla g(x) - \nabla g(y), x - y \rangle \geq 0$.

$$\begin{aligned}
\langle \nabla g(x) - \nabla g(y), x - y \rangle &= L\|x-y\|_2^2 - \langle \nabla f(x) - \nabla f(y), x - y \rangle \\
&\geq L\|x-y\|_2^2 - \|\nabla f(x) - \nabla f(y)\|_2 \cdot \|x-y\|_2, \ (csi) \\
&\geq L\|x-y\|_2^2 - L\|x-y\|_2^2 = 0. \ (smoothness)
\end{aligned}$$

$\square$

It follows that $g(x) = \frac{L}{2}\|x\|_2^2 - f(x)$ is convex.

Smoothness

Fact: f is L-smooth iff $g(x) = \frac{L}{2}\|x\|_2^2 - f(x)$ is convex.

Proof.

$\Leftarrow$) since g is convex we have using def 2:

$$g(y) \geq g(x) + \langle \nabla g(x), y - x \rangle$$

Smoothness

Fact: f is L-smooth iff $g(x) = \frac{L}{2}\|x\|_2^2 - f(x)$ is convex.

Proof.

$\Leftarrow$) since g is convex we have using def 2:

$$g(y) \geq g(x) + \langle \nabla g(x), y - x \rangle$$

$$\Rightarrow \frac{L}{2}\|y\|_2^2 - f(y) \geq \frac{L}{2}\|x\|_2^2 - f(x) + \langle Lx - \nabla f(x), y - x \rangle$$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2$$

$\square$

If f is twice differentiable, then g is also twice differentiable and g convex $\Leftrightarrow \nabla^2 g(x) \succcurlyeq 0 \Leftrightarrow L \cdot I_d - \nabla^2 f(x) \succcurlyeq 0 \Leftrightarrow \nabla^2 f(x) \preccurlyeq L \cdot I_d$

Smoothness: Examples

Example

1. Convex quadratics $x \mapsto x^\top Ax + b^\top x + c$ is smooth for any $L \geq 2\lambda_{\max}(A)$.
2. Logistic : $x \mapsto log(1 + e^{-y\langle a,x\rangle})$
3. Trigonometric : $x \mapsto sin(x), cos(x)$.

Proof.

Exercise ! $\square$

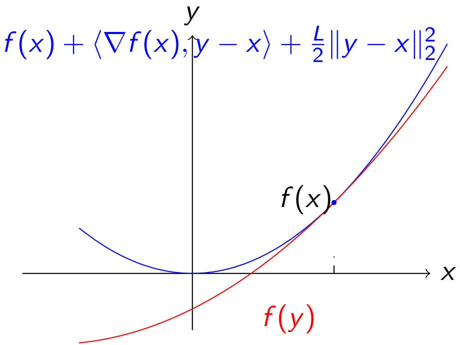
Exercise: Using the fact that $\sigma^2_{\max}(X)\|d\|_2^2 \geq \|X^\top d\|_2^2$, show that

$$f(w) = \frac{1}{2}\|X^\top w - b\|_2^2$$

is $\sigma^2_{\max}(X)$-smooth.

Smoothness: convex example

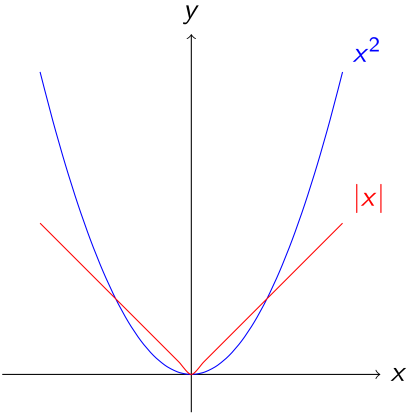
$f(y) = \|y\|_2^2$



$$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2, \quad \forall\, x, y \in \mathbb{R}^d$$

Smoothness: convex counter-example

$f(y) = \|y\|_1$



False : $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2, \quad \forall\, x, y \in \mathbb{R}^d$

Definition

A function f is **strongly convex** with parameter μ if $g(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex.
A function f is convex if it is strongly convex with parameter $\mu = 0$.

$$
\begin{aligned}
g \text{ convex} &\Rightarrow g(y) \geq g(x) + \langle \nabla g(x), y - x \rangle \\
&\Rightarrow f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2
\end{aligned}
\tag{16}
$$

$$
f \text{ is } \mu\text{-strongly convex} \Rightarrow g(x) = f(x) - \frac{\mu}{2}\|x\|_2^2 \text{ is convex}
$$

If f is twice differentiable, then g is also twice differentiable and g convex $\Leftrightarrow \nabla^2 g(x) \succcurlyeq 0 \Leftrightarrow \nabla^2 f(x) - \mu \cdot I_d \succcurlyeq 0 \Leftrightarrow \mu \cdot I_d \preccurlyeq \nabla^2 f(x)$

Recap

Definition

f is μ-strongly convex if

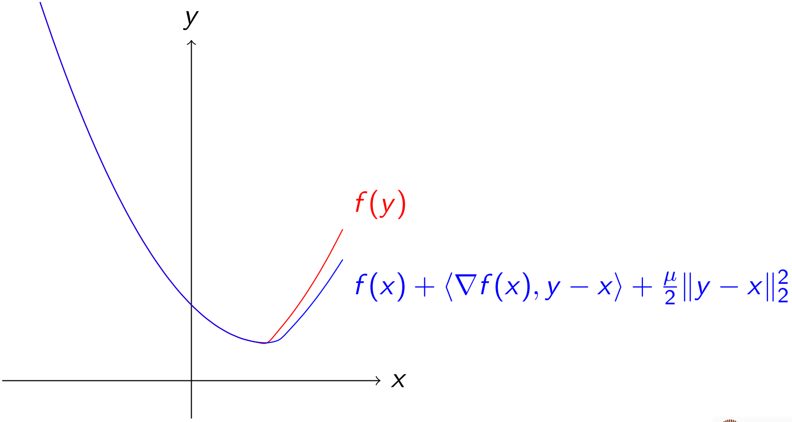
1. $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2$
2. $\mu \cdot I_d \preccurlyeq \nabla^2 f(x)$

If f is μ-strongly convex and L-smooth we have:

$$\mu \cdot I_d \preccurlyeq \nabla^2 f(x) \preccurlyeq L \cdot I_d \tag{17}$$

Strong convexity example

Hinge Loss + L2 : $f(y) = \max(0, 1 - y) + \frac{1}{2}\|y\|_2^2$ is a strongly convex function.



Strong convexity example

1. Using the fact that $\sigma^2_{\min}(X)\|d\|_2^2 \le \|X^\top d\|_2^2$, show that

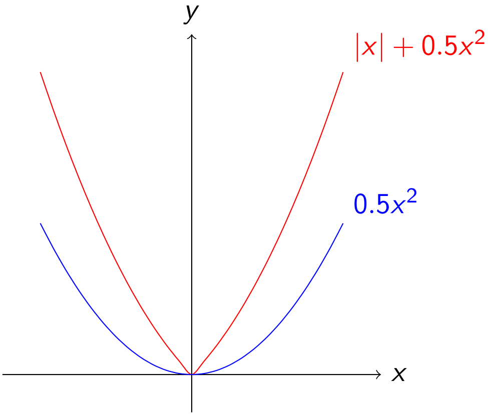
$$f(w) = \frac{1}{2}\|X^\top w - b\|_2^2$$

   is $\sigma^2_{\min}(X)$-strongly convex.

2. $f(x) = x^\top Q x$ is μ-strongly convex with $\mu = 2\lambda_{\min}(Q)$
3. $f(x) = \frac{1}{2}\|x\|_2^2 + \|x\|_1$

Example of strongly convex function and non-smooth

$f(y) = \frac{1}{2}\|x\|_2^2 + \|x\|_1$ is a strongly convex function but non smooth.



Properties of Smooth and (-strongly) convex functions

f convex : $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$.
f L-smooth : $\|\nabla f(y) - \nabla f(x)\|_2 \leq L \cdot \|y - x\|_2$.

Claim: Gradient step guarantees improvement (Gradient descent is a descent algorithm) for smooth functions for small step sizes.

GD: $x^+ = x - \eta \nabla f(x)$, we want to show that $f(x^+) < f(x)$.

Proof.

f is L-smooth $\Rightarrow f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2$ applying the above formular for $y = x^+$ gives :

$$\begin{aligned} f(x^+) &\leq f(x) + \langle \nabla f(x), x^+ - x \rangle + \frac{L}{2}\|x^+ - x\|_2^2 \\ &\leq f(x) - \eta \langle \nabla f(x), \nabla f(x) \rangle + \frac{L\eta^2}{2}\|\nabla f(x)\|_2^2 \end{aligned}$$

Properties of Smooth and (-strongly) convex functions

Proof.

f is L-smooth $\Rightarrow f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2$ applying the above formular for $y = x^+$ gives :

Properties of Smooth and (-strongly) convex functions

Proof.

f is L-smooth $\Rightarrow f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2$ applying the above formular for $y = x^+$ gives :

$$\begin{aligned} f(x^+) &\leq f(x) - \eta\langle \nabla f(x), \nabla f(x) \rangle + \frac{L\eta^2}{2}\|\nabla f(x)\|_2^2 \\ &\leq f(x) - \eta(1 - \frac{\eta L}{2})\|\nabla f(x)\|_2^2 \end{aligned}$$

If η small enough $\eta < \frac{2}{L}$ then $\eta(1 - \frac{\eta L}{2}) > 0$. Typically we choose $\eta = \frac{1}{L}$ so that $x^+ = x - \frac{1}{L}\nabla f(x)$.
Finally we have $f(x^+) \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2 \Rightarrow f(x^+) < f(x)$.
So gradient descent is a descent algorithm. □

A bound on suboptimality of any point

If f is L-smooth :

$$\frac{1}{2L}\|\nabla f(x)\|_2^2 \leq f(x) - f(x^*) \leq \frac{L}{2}\|x - x^*\|_2^2 \tag{18}$$

where x^* is solution to $\min_x f(x)$.
We have $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2 \quad (*)$
since f is L-smooth, taking $y = x, x = x^*$ give us :

A bound on suboptimality of any point

If f is L-smooth :

$$\frac{1}{2L}\|\nabla f(x)\|_2^2 \leq f(x) - f(x^*) \leq \frac{L}{2}\|x - x^*\|_2^2 \tag{18}$$

where x^* is solution to $\min_x f(x)$.
We have $f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2 \quad (*)$
since f is L-smooth, taking $y = x, x = x^*$ give us :

$$f(x) \leq f(x^*) + \langle \nabla f(x^*), y - x^*\rangle + \frac{L}{2}\|y - x^*\|_2^2$$

$$f(x) - f(x^*) \leq \frac{L}{2}\|x - x^*\|_2^2$$

$$f(x^*) \leq f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2$$

Minimizing the quadratic upper bound over y :

$$f(x^*) \leq f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2$$

Minimizing the quadratic upper bound over y :

$$\nabla f(x) + \frac{L}{2}(y - x) = 0$$

$$y = x - \frac{2}{L}\nabla f(x)$$

Now plugging back the value give us

$$f(x) - f(x^*) \geq \frac{1}{2L}\|\nabla f(x)\|_2^2$$

Smoothness: Co-coercivity

If f is L-smooth :

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|_2^2 \tag{19}$$

Proof.

Let define $f_x(z) = f(z) - \langle \nabla f(x), z \rangle$ and $f_y(z) = f(z) - \langle \nabla f(y), z \rangle$. Then

- $z^* = x$ is a minimizer of $f_x(z)$ since $\nabla_z f_x(z) = \nabla f(z) - \nabla f(x) = 0 \Rightarrow z = x$.
- $z^* = y$ is a minimizer of $f_y(z)$.

$$\begin{aligned} f(y) - \big(f(x) + \langle \nabla f(y), y - x \rangle\big) &= f(y) - \langle \nabla f(y), y \rangle - \big(f(x) - \langle \nabla f(x), x \rangle\big) \\ &= f_x(y) - f_x(x) \\ &\geq \frac{1}{2L} \|\nabla f_x(y)\|_2^2 \quad (18) \\ &= \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2 \end{aligned}$$

Smoothness: Co-coercivity

Similarly by flipping roles of x, y:

$$f(x) - \big(f(y) + \langle \nabla f(x), x - y \rangle\big) \geq \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2,$$

adding the two inequalities give the Co-coercivity property.

Strong convexity

recall: f is μ-strongly convex if $g(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex.
If f is μ-strongly convex then,

$$\frac{\mu}{2}\|x - x^*\|_2^2 \leq f(x) - f(x^*) \leq \frac{1}{2\mu}\|\nabla f(x)\|_2^2 \tag{20}$$

Proof.

Suppose f is μ-strongly convex, then
$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2$ taking $y = x, x = x^*$ give us :

Strong convexity

recall: f is μ-strongly convex if $g(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex.
If f is μ-strongly convex then,

$$\frac{\mu}{2}\|x - x^*\|_2^2 \leq f(x) - f(x^*) \leq \frac{1}{2\mu}\|\nabla f(x)\|_2^2 \tag{20}$$

Proof.

Suppose f is μ-strongly convex, then
$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2$ taking $y = x, x = x^*$ give us :

- $f(x) - f(x^*) \geq \frac{\mu}{2}\|x - x^*\|_2^2$
- $f(x^*) \geq \min_y : f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2$

Take the derivative and set to 0: $y = x - \frac{2}{\mu}\nabla f(x)$
$\Rightarrow f(x) - f(x^*) \leq \frac{1}{2\mu}\|\nabla f(x)\|_2^2$ $\square$

Strong Convexity: Coercivity

If f is μ-strongly convex, then

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \cdot \|x - y\|_2^2 \tag{21}$$

Proof.

f is μ-strongly convex if $g(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex.

$$\begin{aligned} g(x) \quad \text{convex} &\Leftrightarrow \nabla g(x) \quad \text{monotone} \\ &\Leftrightarrow \langle \nabla g(x) - \nabla g(y), x - y \rangle \geq 0 \\ &\Leftrightarrow \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \cdot \|x - y\|_2^2 \end{aligned}$$

$\square$

Convergence Rate: Subgradient method and Gradient descent

1. Subgradient method: $x^+ = x - \eta g_x, \quad g_x \in \partial f(x)$
2. Gradient descent for smooth functions
3. Gradient descent for smooth and strongly convex functions

Goal: obtain bounds on sub-optimality of x_k, $f(x_k) - f(x^*)$, where x^* solves $\min_x f(x)$

- Obtain upper-bounds in terms of parameters of the problem
- What does this gap look like as a function of k.

Subgradient method for Lipschitz convex functions

Let f be convex. Assume $\forall\, x, \forall\, g_x \in \partial f(x), \|g_x\| \leq G$.
Subgradient method: $x_{k+1} = x_k - \eta g_k,\ g_k \in \partial f(x_k)$

- Subgradient method is not necessarily a descent method i.e. we can have $f(x_{k+1}) > f(x_k)$

$$\begin{aligned}
\|x_{k+1} - x^*\|_2^2 &= \|x_k - \eta g_k - x^*\|_2^2 \\
&= \|x_k - x^*\|_2^2 - 2\eta\langle g_k, x_k - x^*\rangle + \eta^2\|g_k\|_2^2 \\
&\leq \|x_k - x^*\|_2^2 - 2\eta(f(x_k) - f(x^*)) + \eta^2 G^2
\end{aligned}$$

$$f(x_k) - f(x^*) \leq \frac{1}{2\eta}(\|x_{k+1} - x^*\|_2^2 - \|x_k - x^*\|_2^2) + \frac{\eta}{2}G^2$$

$$\frac{1}{T}\sum_{k=1}^{T}(f(x_k) - f(x^*)) \leq \frac{1}{2\eta T}(\|x_1 - x^*\|_2^2 - \|x_T - x^*\|_2^2) + \frac{\eta}{2}G^2$$

$$f\left(\frac{1}{T}\sum x_k\right) - f(x^*) \leq \frac{1}{T}\sum_{k=1}^{T}(f(x_k) - f(x^*)) \leq \frac{R^2}{2\eta T} + \frac{\eta}{2}G^2$$

where $\|x_1 - x^*\|_2^2 \leq R^2$, this is just saying that we initialize somewhere and we just know that it is not infinitely far away from our solution.

- The best $\eta = \frac{1}{\sqrt{T}}$

Summary of subgradient method: If we plan to run for T iterations best step size $\eta \sim \frac{1}{\sqrt{T}}$.

- Error after T iterations scale like $\sim \frac{1}{\sqrt{T}}$
- To have error ε we need $\sim \frac{1}{\varepsilon^2}$ iterations.

This is a **good news** : subgradient descent works

1. subgradient method produces ε- optimal solutions
2. "dimension free"

Gradient descent for smooth convex functions

Theorem

*Let $f : \mathbb{R}^d \to \mathbb{R}$ be **convex** and L-**smooth** with minimum $f^* = f(x^*)$ and let x_k be defined by : $x_{k+1} = x_k - \eta \nabla f(x_k)$ for $0 < \eta < \frac{2}{L}$. Then it holds that*

$$f(x_k) - f^* \leq \frac{2(f(x_0) - f^*)\|x_0 - x^*\|_2^2}{2\|x_0 - x^*\|_2^2 + k\eta(2 - L\eta)(f(x_0) - f^*)} \sim \mathcal{O}(\frac{1}{k}) \qquad (22)$$

Proof

We denote by $r_k = \|x_k - x^*\|_2^2$ and estimate

$$\begin{aligned} r_{k+1} &= \|x_{k+1} - x^*\|_2^2 \\ &= r_k - 2\eta\langle \nabla f(x_k), x_k - x^* \rangle + \eta^2 \|\nabla f(x_k)\|_2^2 \\ &= r_k - \frac{2\eta}{L}\|\nabla f(x_k) - \nabla f(x^*)\|_2^2 + \eta^2 \|\nabla f(x_k)\|_2^2 \\ &= r_k - \eta(\frac{2}{L} - \eta)\|\nabla f(x_k)\|_2^2 \end{aligned}$$

We see that $r_k \leq r_0$ and we get (denoting $w = \eta(1 - \frac{L}{2}\eta)$)

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 \\ &= f(x_k) - w\|\nabla f(x_k)\|_2^2 \end{aligned}$$

Proof

We further abbreviate $\Delta_k = f(x_k) - f^*$ and get by convexity of f and Cauchy-Schwarz

$$\Delta_k \leq \langle \nabla f(x_k), x_k - x^* \rangle \leq r_k \cdot \|\nabla f(x_k)\|_2 \leq r_0 \cdot \|\nabla f(x_k)\|_2$$

Together with the above we obtain $f(x_{k+1}) \leq f(x_k) - \frac{w}{r_0}\Delta_k^2 \Rightarrow \Delta_{k+1} \leq \Delta_k - \frac{w}{r_0}\Delta_k^2 = \Delta_k\left(1 - \frac{w}{r_0}\Delta_k\right)$ which implies $\Delta_{k+1} \leq \Delta_k$ and can be rearranged to

$$\frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{w}{r_0}\frac{\Delta_k}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{w}{r_0} \geq \cdots \geq \frac{1}{\Delta_0} + \frac{w}{r_0}(k+1)$$

this finally gives

$$\Delta_k \leq \frac{1}{\frac{1}{\Delta_0} + \frac{w}{r_0}k} = \frac{\Delta_0 r_0}{r_0 + \Delta_0 wk}$$

To get a clearer bound, we optimize the right hand side over the step-size η and get :

Corollary

*Let $f : \mathbb{R}^d \to \mathbb{R}$ be **convex** and L-**smooth** with minimum $f^* = f(x^*)$ and let x_k be defined by : $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$. Then it holds that*

$$f(x_k) - f^* \leq \frac{2L\|x_0 - x^*\|_2^2}{k+4} \sim \mathcal{O}(\frac{1}{k}) \tag{23}$$

We want to make $\eta(2 - L\eta)$ as large as possible, and this is the case for $\eta^* = \frac{1}{L}$. Then $\eta^*(2 - L\eta) = \frac{1}{L}$. Furthermore, use L-smoothness and $\nabla f(x^*)$ to estimate $f(x_0) - f^* \leq \frac{L}{2}\|x_0 - x^*\|_2^2$. This simplifies the upper bound from previous Theorem to $\frac{2L\|x_0 - x^*\|_2^2}{k+4}$.

Gradient descent for smooth and strongly convex functions

Assume f is μ-strongly convex and L-smooth i.e. $f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex. Also this is $(L-\mu)$ smooth (i.e. $\frac{(L-\mu)}{2}\|x\|_2^2 - f(x)$ is convex.)
Co-coercivity of $f(x) - \frac{\mu}{2}\|x\|_2^2$ implies:

$$
\begin{aligned}
\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu\|x-y\|_2^2 + \frac{1}{L-\mu}\|\nabla f(x) - \nabla f(y)\|_2^2 + \\
\frac{\mu^2}{L-\mu}\|x-y\|_2^2 - \frac{2\mu}{L-\mu}\langle \nabla f(x) - \nabla f(y), x - y \rangle
\end{aligned}
$$

simplify, we find:

Gradient descent for smooth and strongly convex functions

Assume f is μ-strongly convex and L-smooth i.e. $f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex.
Also this is $(L-\mu)$ smooth (i.e. $\frac{(L-\mu)}{2}\|x\|_2^2 - f(x)$ is convex.)
Co-coercivity of $f(x) - \frac{\mu}{2}\|x\|_2^2$ implies:

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu\|x-y\|_2^2 + \frac{1}{L-\mu}\|\nabla f(x) - \nabla f(y)\|_2^2 + \frac{\mu^2}{L-\mu}\|x-y\|_2^2 - \frac{2\mu}{L-\mu}\langle \nabla f(x) - \nabla f(y), x - y \rangle$$

simplify, we find:

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \underbrace{\frac{\mu}{L+\mu}\|x-y\|_2^2}_{Coercivity} + \underbrace{\frac{1}{L+\mu}\|\nabla f(x) - \nabla f(y)\|_2^2}_{Co-coercivity}$$

Theorem

*Let $f : \mathbb{R}^d \to \mathbb{R}$ be μ-**strongly convex** and L-**smooth** with minimum $f^* = f(x^*)$ and let x_k be defined by : $x_{k+1} = x_k - \eta \nabla f(x_k)$ for $0 < \eta < \frac{2}{L}$. Then it holds that*

$$\|x_k - x^*\|_2^2 \leq \left(1 - \frac{2\eta\mu L}{\mu + L}\right)^k \cdot \|x_0 - x^*\|_2^2 \qquad (24)$$

for $0 < \eta \leq 2/(L + \mu)$. The right hand side is minimal for the step-size $\eta = \frac{2}{L+\mu}$ and in this case we get

$$\|x_k - x^*\|_2 \leq \left(\frac{Q-1}{Q+1}\right)^k \cdot \|x_0 - x^*\|_2$$

$$f(x_k) - f^* \leq \frac{L}{2}\left(\frac{Q-1}{Q+1}\right)^{2k} \cdot \|x_0 - x^*\|_2^2$$

with condition number $Q = \frac{L}{\mu}$.

Acknowledgements

Thanks for your attention!