1 Convexity

Cauchy-Schwarz $\|\mathbf{u}^\top \mathbf{u}\| \leq |\mathbf{u}||\mathbf{v}| \rightarrow \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} \in [-1,1] \sim \cos\alpha$

Definition of convexity A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex if (i) $\mathbf{dom}(f)$ is a convex set and (ii) $\forall \mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$ and $\lambda \in [0,1]$ we have

$$\underbrace{f\left(\lambda x + (1-\lambda)y\right)}_{(1)} \leq \underbrace{\lambda f(x) + (1-\lambda) f(y)}_{(2)}$$

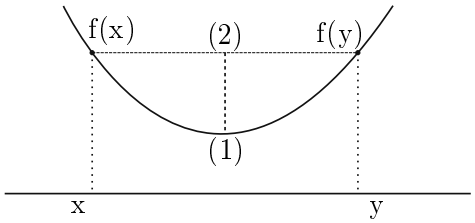


Figure 1: Any point on the line is above f

Epigraph While the *graph* of a function is the line drawn by its expression, the *epigraph* is the set of points that lie above the graph (the "content" of the graph).

Jensen's inequality The above definition is valid for any number of points in $\mathbf{dom}(f)$. Any "middle point" will be in-between them, and always above f. Formally: Let f be convex, and $x_1, \ldots, x_m \in \mathbf{dom}(f)$, $\lambda_1, \ldots, \lambda_m \in \mathbb{R}_+$ such that $\sum_{i=1}^m \lambda_i = 1$. Then

$$f\left(\sum_{i=1}^m \lambda_i \mathbf{x}_i\right) \leq \sum_{i=1}^m \lambda_i f(\mathbf{x}_i).$$

Convex is continuous Let f be convex, and suppose that $\mathbf{dom}(f)$ is open. Then f is continuous. is continuous.

Differentiable Graph of the affine function $f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$ is a tangent hyperplane to the graph of f at $(\mathbf{x}, f(\mathbf{x})$

First-order characterization Suppose $\mathbf{dom}(f)$ is open and $f(\mathbf{x})$ is differentiable, in particular the gradient exists at every point $\mathbf{x} \in \mathbf{dom}(f)$. Then f is convex if and only if $\mathbf{dom}(f)$ is convex and

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$

This means that f is above all its tangent hyperplanes.

Second-order Characterization Suppose that $\mathbf{dom}(f)$ is open, and that the Hessian (double derivatives) of f exists at every point $\mathbf{x} \in \mathbf{dom}(f)$ and is symmetric. Then f is convex if and only if $\mathbf{dom}(f)$ is convex, and for all $\mathbf{x} \in \mathbf{dom}(f)$ we have

$$\nabla^2 f(\mathbf{x}) \succeq 0^1$$

Operations Multiplication by real constant and addition between convex is convex (not everywhere, only on $\bigcap_{i=1}^m \mathbf{dom}(f_i)$). The composition works the following: Let f be a convex function with $\mathbf{dom}(f) \subseteq \mathbb{R}^d$, $g : \mathbb{R}^m \rightarrow \mathbb{R}^d$ an affine function, meaning that $g(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$, for some matrix $A \in \mathbb{R}^{d \times m}$ and some vector $\mathbf{b} \in \mathbb{R}^d$. Then the function $f \circ g$ (that maps $\mathbf{x}$ to $f(A\mathbf{x} + \mathbf{b})$ is convex on $\mathbf{dom}(f \circ g) := \{\mathbf{x} \in \mathbb{R}^m : g(\mathbf{x}) \in \mathbf{dom}(f)\}$

Mimima $\mathbf{x}$ is a local minimum if $\exists \epsilon > 0$ with

$$f(\mathbf{x}) \leq f(\mathbf{y})$$

For all $\mathbf{y} \in \mathbf{dom}(f)$ satisfying $\|\mathbf{y} - \mathbf{x}\| < \epsilon$. If $\mathbf{x}^\star$ is the local minimum of a convex function, then it's a global minimum (meaning $f(\mathbf{x}^\star) \leq f(\mathbf{y})$ for all $\mathbf{y} \in \mathbf{dom}(f)$. Similarly, for f convex and differentiable over an open domain, then if $\nabla f(\mathbf{x}) = \mathbf{0}$ then it's a global minimum (it's called a *critical point*).

Strictly convex Same definition as the convexity, with a strict inequality. So the open

[1] positive semidefinite

segment connecting any two points of the graph will be *strictly* above the graph. This leads to the following lemma:

Lemma. *Suppose that* $\mathbf{dom}(f)$ *is open and that* f *is twice continuously differentiable. If the Hessian* $\nabla^2 f(\mathbf{x}) \succeq \mathbf{0}$ *for every* $\mathbf{x} \in \mathbf{dom}(f)$*, then* f *is strictly convex.*

Constrained Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be convex, and let $X \subseteq \mathbf{dom}(f)$ be a convex set. A point $\mathbf{x} \in X$ is a *minimizer of* f *over* X if $f(\mathbf{x}) \leq f(\mathbf{y})\ \forall \mathbf{y} \in X$. It follows that for $f : \mathbf{dom}(f) \to \mathbb{R}$ convex and differentiable oven an open domain $\mathbf{dom}(f) \subseteq \mathbb{R}^d$, and for $X \subseteq$ a convex set, then the point $\mathbf{x}^\star \in X$ is a minimizer of f over X iff

$$\nabla f(\mathbf{x}^\star)^\top (\mathbf{x} - \mathbf{x}^\star) \geq 0\ \forall \mathbf{x} \in X$$

α-sublevel $f^{\leq \alpha} := \{\mathbf{x} \in \mathbb{R}^d : f(\mathbf{x}) \leq \alpha$. This represents the values of the domain of f for which the value of f is below a threshold

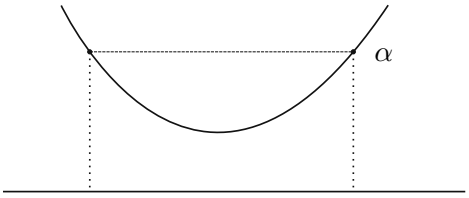


Figure 2: Only what is between vertical bars is in the sublevel

Weierstrass theorem Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, and suppose there is a nonempty and bounded sublevel set $f^{\leq \alpha}$. Then f has a global minimum. Some function (such as e^x) don't have a minimum (in the exponential case, because the subvlevel is not bounded.

2 Gradient descent

Gradient descent Goal: get near to a minimum $x^\star$ (not necessarily unique), meaning close to the optimal value $f(x^\star)$. For that, we look for $\mathbf{x} \in \mathbb{R}^d$ such that

$$f(\mathbf{x}) - f(\mathbf{x}^\star) \leq \epsilon$$

Iterative algorithm: For that purpose, we start from $\mathbf{x}_0 \in \mathbb{R}^d$, and then iterate:

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t)$$

For time steps t=0,1,..., and step size $\gamma \geq 0$

Bound to error (vanilla) It's useful to bound the error $(f(\mathbf{x}_t) - f(\mathbf{x}^\star))$. Using the notation $\mathbf{g}_t := \nabla f(\mathbf{x}_t)$ (for gradient descent, $\mathbf{g}_t = \frac{\mathbf{x}_t - \mathbf{x}_{t+1}}{\gamma}$), and applying the following steps:

1. Apply $2\mathbf{v}^\top \mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2$
2. Sum over for the first T iterations
3. Use first-order characterization with $\mathbf{x} = \mathbf{x}_t, \mathbf{y} = \mathbf{x}^\star$.

We obtain the upper bound for the average error:

$$\sum_{t=0}^{T-1} (f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq \frac{\gamma}{2} \sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2$$

Lipschitz

Theorem. *A function* $f : \mathbf{dom}(f) \to \mathbb{R}^m$ *is B-Lipschitz if*

$$\|f(\mathbf{x}) - f(\mathbf{y})\| \leq B\|\mathbf{x} - \mathbf{y}\|$$

for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$

Assume all gradients of f bounded in norm (not always cool, e.g. discards x^2). The following theorem holds:

Theorem. *Let* $f : \mathbb{R}^d \to \mathbb{R}$ *be convex and differentiable, and suppose* $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$*, and* $\|\nabla f(\mathbf{x})\| \leq B$ *for all* $\mathbf{x}$*. Choosing the step size*

$$\gamma := \frac{R}{B\sqrt{T}}$$

then gradient descent yields

$$\frac{1}{T}\sum_{t=0}^{T-1} f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{RB}{\sqrt{T}}$$

This relates T and ϵ !

$$T \geq \frac{R^2B^2}{\epsilon^2} \Rightarrow \text{average error } \leq \frac{RB}{\sqrt{T}} \leq \epsilon$$

Note: same for subgradient descent.

Smoothness A function is smooth if it is "Not too curved". That is, if given a point $\mathbf{x} \in X$, all other points smaller than the linearisation + a quadratic term. Formally, let $f : \mathbf{dom}(f) \to \mathbb{R}$ be differentiable, $X \subset \mathbf{dom}(f)$, $L \in \mathbb{R}_+$. f is called smooth (with parameter L over X if

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

$\forall \mathbf{x}, \mathbf{y} \in X$.

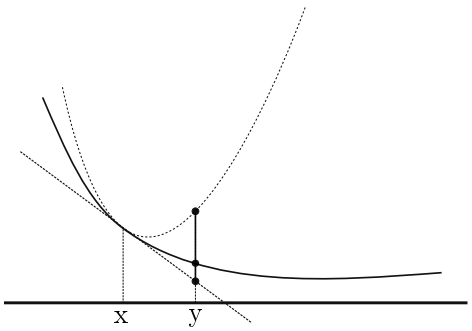


Figure 3: Every $\mathbf{y}$ is below the linearisation+quadratic

(unsure) The smoothness of least squares ($f(x) = \frac{1}{2n}\|A\mathbf{x}+\mathbf{b}\|$) is given by $\max \frac{\|A^\top \cdot A\|}{n}$. That is the maximum of the eigenvalues of $A^\top \cdot A$ divided by the number of rows of A.

Operations on Smoothness Addition of smooth function is still smooth: $f_1, f_2, \ldots, f_m$ smooth with parameters $L_1, L_2, \ldots, L_m$ and let $\lambda_1, \lambda_2, \ldots, \lambda_m \in \mathbb{R}_+$. Then $f := \sum_{i=1}^m \lambda_i L_i$ is smooth, with parameter $\sum_{i=1}^m \lambda_i L_i$.

Also, about composition: f L-smooth, and let $g(\mathbf{x}) = A\mathbf{x}+\mathbf{b}$ for $A \in \mathbb{R}^{d\times m}$, $\mathbf{b} \in \mathbb{R}^d$. Then $f \circ g$ is smooth with parameter $L\|A\|^2$ (where $\|A\|$ is the spectral norm of A).

Smooth VS Lipschitz

1. Bounded gradients $\iff$ Lipschitz continuity of f
2. Smoothness $\iff$ Lipschitz continuity of ∇f

For $f : \mathbb{R}^d \to \mathbb{R}$, convex and differentiable, the following are equivalent:

- f is smooth with parameter L
- $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x}-\mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

Sufficient decrease $f : \mathbb{R}^d \to \mathbb{R}$ differentiable, L-smooth, step size $\gamma := \frac{1}{L}$. Then gradient descent satisfies

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 \quad t \geq 0$$

This is a strong statement. This implies that the steps always get better (second term is always positive)

Smooth & convex f convex, differentiable, with global minimum, L-smooth, step size $\frac{1}{L}$. Gradient descent yields

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \; T > 0$$

This relates T and ϵ !

$$T > \frac{R^2L}{2\epsilon} \Rightarrow \text{error} \leq \frac{L}{2T} \leq \epsilon$$

Strongly convex Up until now, error decreased in T or $\sqrt{T}$. Good, but not ideal.

Strongly convex functions converge exponentially. For $f : \mathbf{dom}(f) \to \mathbb{R}$ differentiable function, $X \subset \mathbf{dom}(f)$ convex and $\mu \in \mathbb{R}_+,\ \mu > 0$. Then f is strongly convex (with parameter μ) over X if

$$f(\mathbf{y}) \geq \overbrace{f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})}^{\text{Linearisation at } x} + \underbrace{\frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2}_{\text{Quadratic term}} \quad \forall \mathbf{x}, \mathbf{y} \in X$$

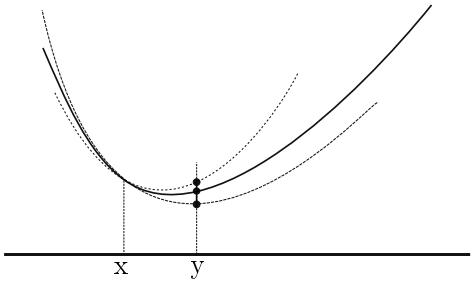


Figure 4: Function is between smooth (above) and strongly convex (below)

Smooth & Strongly convex We start from the vanilla analysis, and use stronger lower bound on left-hand side (coming from strong convexity). That is, we use that

$$\nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^\star) \geq f(\mathbf{x}_t) - f(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2$$

and some rewriting to obtain the final bound:

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq & 2\gamma \left(f(\mathbf{x}^\star) - f(\mathbf{x}_t)\right) \\ & + \gamma^2 \|\nabla f(\mathbf{x}_t)\| \\ & + (1 - \mu\gamma)\|\mathbf{x}_t - \mathbf{x}^\star\|^2\end{aligned}$$

In other words: *Squared distance to $\mathbf{x}^\star$ goes down by a constant factor, up to some "noise".* This leads to the following theorem:

Theorem. *Let $f : \mathbb{R}^d \to \mathbb{R}$, differentiable, with a global minimum $x^\star$. Suppose it's smooth with parameter L and strongly convex with parameter $\mu > 0$. With $\gamma := \frac{1}{L}$, gradient descent with arbitrary $\mathbf{x}_0$ satisfies the following 2 properties:*

- *Squared distance to $\mathbf{x}^\star$ are geometrically decreasing:*

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \left(1 - \frac{\mu}{L}\right) \|\mathbf{x}_t - \mathbf{x}^\star\|^2,\ t \geq 0$$

- *The absolute error after T iterations is exponentially small in T :*

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2}\left(1 - \frac{\mu}{L})\right)^T \|\mathbf{x}_0 - \mathbf{x}^\star\|^2$$

3 Projected Gradient Descent

Constrained Optimization We want to minimize $f(\mathbf{x})$, but being subject to $\mathbf{x} \in X$. 2 ways of solving this: Either use projected gradient descent, or transform it into an *unconstrained* problem.

Projected Gradient Descent Idea: project onto X after every step:

$$\Pi_X(\mathbf{y}) := \arg\min_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{y}\|$$

Idea: compute first what would be your next step ($\mathbf{y}_{t+1}$) and then project it on your actual next step.

$$\begin{aligned}\mathbf{y}_{t+1} &:= \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t), \\ \mathbf{x}_{t+1} &:= \Pi_X(\mathbf{y}_{t+1}) := \arg\min_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{y}_{t+1}\|^2\end{aligned}$$

Properties Let $X \in \mathbb{R}^d$ be closed and convex, $\mathbf{x} \in X, \mathbf{y} \in \mathbb{R}^d$. Then

- $(\mathbf{x} - \Pi_X(\mathbf{y}))^\top (\mathbf{y} - \Pi_X(\mathbf{y})) \leq 0$
- $\|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2$

First property means that for any point $\mathbf{y}$, and any point $\mathbf{x}$ in the domain, then the angle

formed at the projection of $\mathbf{y}$ (in X) is greater than 90° (negative inner product).

Second property is the triangle inequality: distance x-projection plus projection-$\mathbf{y}$ is bigger than distance $\mathbf{x} - \mathbf{y}$. Inequality is written in the other round because of the squared values. If you remove the square, then inequality reverts.

Last cool property: The expected number of steps in various scenarios (Lipschitz,...), as listed in appendix Table 1, remains unchanged!

Equivalent proofs of projected The last property abode needs to be proved (not here). However, here are some insights: We now only need the function to be continuous over X, and not anymore $\mathbb{R}^d$. Then, in the proofs we replace $\mathbf{x}_{t+1}$ by $\mathbf{y}_{t+1}$, and massage it with the facts above.

Smooth and strongly convex over X

Theorem. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. Let $D \subseteq \mathbb{R}^d$ be a normal nonempty closed and convex set and suppose that f is smooth over X with parameter L and strongly convex over X with parameter $\mu > 0$. Choosing $\gamma := \frac{1}{L}$, **projected** gradient descent with arbitrary $\mathbf{x}_0$ satisfies the following two properties:*

1. *Squared distance to $\mathbf{x}^\star$ are geometrically decreasing:*

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \left(1 - \frac{\mu}{L}\right) \|\mathbf{x}_t - \mathbf{x}^\star\|^2, \ t \geq 0$$

2. *The absolute error after T iterations is exponentially small in T:*

$$\begin{aligned} &f(\mathbf{x}_T) - f(\mathbf{x}^\star) \\ &\leq \|\nabla f(\mathbf{x}^\star)\| \left(1 - \tfrac{\mu}{L}\right)^{T/2} \|\mathbf{x}_0 - \mathbf{x}^\star\| \\ &+ \tfrac{L}{2} \left(1 - \tfrac{\mu}{L}\right)^{\top} \|\mathbf{x}_0 - \mathbf{x}^\star\| \end{aligned}$$

Projection step Computing the projection is far from obvious. But for some relevant cases, it can be efficiently solved:

- Projecting onto an affine subspace (leads to system of linear equations)
- Projecting onto an Euclidean ball with center $\mathbf{c}$ (scale vector $\mathbf{y} - \mathbf{c}$). Compute $y = (\mathbf{y} - \mathbf{c}) \cdot \frac{R}{\|\mathbf{x}\|}$.
- Projecting onto ℓ_1-balls (needed in Lasso). Though, we restrict to the ball being centered at $\mathbf{0}$, defined as

$$B_1(R) = \left\{ \mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_1 = \sum_{i=1}^{d} |x_i| \leq R \right\}$$

  B_1 is a *cross polytope* ($2d$ facets, 2^d vertices). The projection can be computed in $\mathcal{O}(d \log d)$, and even improved to $\mathcal{O}(d)$

4 Proximal and Subgradient Descent

Composite optimization problems Consider $f(\mathbf{x}) := g(\mathbf{x}) + h(\mathbf{x})$, with g is "nice" and h is a "simple" additional term (but not nice). Important case when h is not differentiable.

Proximal gradient Classical gradient step:

$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{y}} g(\mathbf{x}_t)^\top (\mathbf{y} - \mathbf{x}_t) + \frac{1}{2\gamma} \|\mathbf{y} - \mathbf{x}_t\|^2$$

For $f = g + h$ keep the same for g and add h unmodified:

$$\begin{aligned} \mathbf{x}_{t+1} &= \arg\min_{\mathbf{y}} g(\mathbf{x}_t)^\top (\mathbf{y} - \mathbf{x}_t) + \tfrac{1}{2\gamma} \|\mathbf{y} - \mathbf{x}_t\|^2 \\ &= \arg\min_{\mathbf{y}} \tfrac{1}{2\gamma} \|\mathbf{y} - (\mathbf{x}_t - \gamma \nabla g(\mathbf{x}_t))\|^2 + h(\mathbf{y}) \end{aligned}$$

Proximal mapping One iteration so proximal gradient descent is

$$\mathbf{x}_{t+1} := \text{prox}_{h,\gamma}(\mathbf{x} - \gamma \nabla g(\mathbf{x}_t))$$

where we define, for a given h and a $\gamma > 0$:

$$\text{prox}_{h,\gamma}(\mathbf{z}) := \arg\min_{\mathbf{y}} \left\{ \frac{1}{2\gamma} \|\mathbf{y} - \mathbf{z}^2 + h(\mathbf{y}\| \right\}$$

Generalized gradient The above allows us to describe the generalized gradient as

$$G_{h,\gamma} := \frac{1}{\gamma} \left(\mathbf{x} - \text{prox}_{h,\gamma}(\mathbf{x} - \gamma \nabla g(\mathbf{x})) \right))$$

and thus rewrite the update step as

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma G_\gamma(\mathbf{x}_t)$$

Generalization Proximal is generalization of gradient descent. If h is 0, it's gradient descent, if it's ι_X, it's the projected gradient descend. With

$$\iota_X(\mathbf{x}) := \begin{cases} 0 & \text{if } \mathbf{x} \in X \\ +\infty & \text{otherwise} \end{cases}$$

Proximal mapping becomes, with ι_X it becomes $\text{prox}_{h,\gamma} = \arg\min_{\mathbf{y} \in X} \|\mathbf{y} - \mathbf{z}\|^2$

Convergence $\mathcal{O}(\frac{1}{\epsilon})$ for smooth (same as vanilla), and if also strongly convex $\mathcal{O}(\log(\frac{1}{\epsilon}))$

Subgradient $\mathbf{g} \in \mathbb{R}^d$ is a subgradient of f at $\mathbf{x}$ if

$$f(\mathbf{y}) \geq \mathbf{g}^\top (\mathbf{y} - \mathbf{x}) \; \forall \mathbf{y} \in \mathbf{dom}(f)$$

Subdifferential $\partial f(\mathbf{x}) \subseteq \mathbb{R}^d$ is the set of subgradients of f at $\mathbf{x}$.

Lemma. *If $f : \mathbf{dom}(f) \to \mathbb{R}$ is differentiable at $\mathbf{x} \in \mathbf{dom}(f)$, then $\partial f(\mathbf{x}) \subseteq \{\nabla f(\mathbf{x})\}$*

Subgradient and convexity "Convex = subgradients everywhere".

Lemma. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ is convex if and only if $\mathbf{dom}(f)$ is convex and $\partial f(\mathbf{x}) \neq \emptyset$ $\forall \mathbf{x} \in \mathbf{dom}(f)$*

Convex and Lipschitz

Lemma. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ convex, $\mathbf{dom}(f)$ open, $B \in \mathbb{R}_+$. Then the following are equivalent:*

- *$\|\mathbf{x} \leq B\|$ for all $\mathbf{x} \in \mathbf{dom}(f)$ and all $\mathbf{g} \in \partial f(\mathbf{x})$.*
- *$|f(\mathbf{x}) - f(\mathbf{y})| \leq B\|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$*

Optimility condition

Lemma. *Suppose $f : \mathbf{dom}(f) \to \mathbb{R}$ and $\mathbf{x} \in \mathbf{dom}(f)$. If $\mathbf{0} \in \partial f(x)$, then $\mathbf{x}$ is a global minimum*

Differentiability

Theorem. *A convex function $f : \mathbf{dom}(f) \to \mathbb{R}$ is differentiable almost everywhere.*

Meaning: set of points where f is non-differentiable has measure 0 (no volume), and for all $\mathbf{x} \in \mathbf{dom}(f)$ and all $\epsilon > 0$, there is a point $\mathbf{x}'$ such that $\|\mathbf{x} - \mathbf{x}'\| < \epsilon$ and f is differentiable at $\mathbf{x}'$. This is a problem for gradient descent. Min can be non-differentiable, and non-differentiable points can be requested during descent.

Subgradient descent To solve this: use subgradient descent. Choose starting point $\mathbf{x}_0 \in \mathbb{R}^d$. Then for each times $t = 0, 1, \ldots$ and stepsizes $\gamma_t \geq 0$ (not necessarily fixed) :

$$\text{Let } \mathbf{g}_t \in \partial f(\mathbf{x}_t)$$
$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t \mathbf{g}_t$$

Optimality Many convergence rates. Can we always improve? No.

Theorem (Nesterov). *For any $T \leq d - 1$ and starting point $\mathbf{x}_0$, there is a function f in the problem class of B-Lipschitz function over $\mathbb{R}^d$, such that any (sub)gradient method has an objective error at least*

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \geq \frac{RB}{2(1 + \sqrt{T+1})}$$

Strongly convex Definition is similar that for gradient, with the use of subgradient.

$$f(\mathbf{y}) \geq \overbrace{f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y}-\mathbf{x})}^{\text{Linearisation at } x} + \underbrace{\frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2}_{\text{Quadratic term}}$$

$\forall \mathbf{x}, \mathbf{y} \in \mathbf{dom}(f), \forall \mathbf{g} \in \partial f(\mathbf{x})$. Alternatively:

Lemma. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be convex, $\mathbf{dom}(f)$ open, $\mu \in \mathbb{R}_+^*$. f is strongly convex with parameter μ if and only if $f_\mu : \mathbf{dom}(f) \to \mathbb{R}$ defined by*

$$f_\mu(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2, \ \mathbf{x} \in \mathbf{dom}(f)$$

is convex

Tame strong convexity Over $\mathbb{R}^d$, strong convexity and subgradients contradict each other. With strong convexity, gradients will explode to infinity.

Theorem. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be strongly convex with parameter $\mu > 0$, and let $\mathbf{x}^\star$ be the unique global minimum of f. With decreasing step size*

$$\gamma_t := \frac{2}{\mu(t+1)}, \ t > 0$$

subgradient descent yields

$$\underbrace{f\left(\frac{2}{T(T+1)}\sum_{t=1}^{T} t \cdot \mathbf{x}_t\right)}_{\text{convex combination of iterates}} - f(\mathbf{x}^\star) \leq \frac{2B^2}{\mu(T+1)}$$

where $B = \max_{t=1}^T \|\mathbf{g}_t\|$

Strong convexity

5 Stochastic Gradient Descent

Idea Many objective functions are sum structured

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

As for example, f_i is the cost function of i-th observation. But doing so on very large datasets (e.g. Imagenet, $n \simeq 14M$). So we only do on a subset.

The algorithm As always, get a starting point $\mathbf{x}_0 \in \mathbb{R}^d$, then sample $i \in [n]$ uniformly at random, and compute as before:

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t \nabla f_i(\mathbf{x}_t)$$

So we only update with the gradient of f_i instead of the full gradient. The vector $\mathbf{g}_t := \nabla f_i(\mathbf{x}_t)$ is called a stochastic gradient.

Analysis We can't use the same as vanilla, because convexity $\big(f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)$ doesn't hold anymore$\big)$. But cool thing! $\mathbf{g}_t$ is an unbiased estimate of the full gradient:

$$\mathbb{E}[\mathbf{g}_t | \mathbf{x}_t = \mathbf{x}] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}) = \nabla f(\mathbf{x})$$

So the convexity inequality holds in expectation.

Holds in expectation For any fixed $\mathbf{x}$, linearity of conditional expectations yields

$$\begin{aligned}\mathbb{E}\left[\mathbf{g}_t(\mathbf{x} - \mathbf{x}^\star) | \mathbf{x}_t = \mathbf{x}\right] &= \mathbb{E}\left[\mathbf{g}_t | \mathbf{x}_t = \mathbf{x}\right]^\top (\mathbf{x} - \mathbf{x}^\star) \\ &= \nabla f(\mathbf{x})^\top (\mathbf{x} - \mathbf{x}^\star)\end{aligned}$$

Bounded stochastic gradient

Theorem. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable, $\mathbf{x}^\star$ a global minimum. Furthermore, suppose that $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$ and that $\mathbb{E}\left[\|\mathbf{g}_t\|^2\right] \leq B^2$ for all t. Choosing the constant stepsize*

$$\gamma := \frac{R}{B\sqrt{T}}$$

stochastic gradient descent yields

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[f(\mathbf{x}_t)\right] - f(\mathbf{x}^\star) \leq \frac{RB}{\sqrt{T}}$$

Convergence rate: SGD vs GD For GD: we assumed $\|\nabla \frac{f}{\mathbf{x})}\|^2 \leq B_{GD}^2$, leading to

$$\|\frac{1}{n}\sum_i \nabla f_i(\mathbf{x})\|^2 \leq B_{GD}^2$$

As for SGD, assuming the same for the expected squared norms of our stochastic gradients, now called B_{SGD}^2 :

$$\frac{1}{n}\sum_i \|\nabla f_i(\mathbf{x})\|^2 \leq B_{SGD}^2$$

So GD can me better, but often comparable. Very similar if larger mini-batches are used:

$$\underbrace{\|\frac{1}{n}\sum_i \nabla f_i(\mathbf{x})\|^2}_{\approx B_{GD}^2} \leq \underbrace{\frac{1}{n}\sum_i \|\nabla f_i(\mathbf{x})\|^2}_{\approx B_{SGD}^2}$$

Strong convexity

Theorem. *For a strongly convex f, and with decreasing step size*

$$\gamma_t := \frac{2}{\mu(t+1)}$$

stochastic gradient descent yields

$$\mathbb{E}\left[f\left(\frac{2}{T(T+1)}\sum_{t=1}^{T} t\cdot \mathbf{x}_t\right) - f(\mathbf{x}^\star)\right] \leq \frac{2B^2}{\mu(T+1)}$$

Mini-Batch Instead of using a single element f_i, use an average of several of them

$$\tilde{\mathbf{g}}_t := \frac{1}{m}\sum_{j=1}^{m} \mathbf{g}_t^j$$

At the extremes ($m = 1$ or $m = n$), we fall back to full gradient descent or SGD. But now computation can be naively parallelized. Also, by taking an average of many independent random variables reduce the variance. With larger size of the mini-batch m, $\tilde{\mathbf{g}}_t$ will be closer to the true gradient, in expectation:

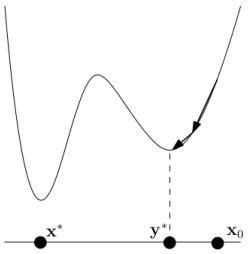
$$Var(\tilde{\mathbf{g}}_t) = \mathbb{E}\left[\|\tilde{\mathbf{g}}_t - \nabla f(\mathbf{x}_t)\|^2\right] \leq \frac{B^2}{m}$$

Stochastic Subgradient Descent For problems not necessarily differentiable. Quite the same, but we use a subgradient for one sample. We have $\mathbf{g}_t \in \partial f_i(\mathbf{x}_t)$. The rest is as before. This yields an unbiased estimate of the subgradient. We have a convergence in $\mathcal{O}(\frac{1}{\epsilon^2})$ by using the subgradient property.

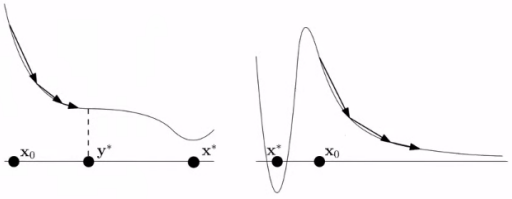
Constrained optimization By slapping a projection back to X onto every step, we obtain projected SGD, with the same convergence rate as above., we obtain projected SGD, with the same convergence rate as above.

6 Non-convex Optimization

Gradient descent Main problem, is that



we may get stuck in a local minimum, or in a saddle point, or run off to infinity,...



Convave f is concave is $-f$ is convex (for all $\mathbf{x}$, f is below the tangent at x)

Bounded Hessians

Lemma. *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be twice differentiable, with $X \subset \mathbf{dom}(f)$ a convex set, and $\|\nabla^2 f(\mathbf{x})\| \leq L$ for all $\mathbf{x} \in X$, where $\|\cdot\|$ is the spectral norm. Then f is smooth with parameter L over X.*

So bounded Hessians ⇒ smooth. The opposite (smooth ⇒ bounded Hessians) is true over any open convex set X.

Convergence We sadly can't prove that we will converge to $x^\star$. But we can show that we will converge to a null gradient, at the same rate that we converge to $x^\star$ in the convex case (for $t \to \infty$)

Smooth (but not necessarily convex). For f smooth with parameter L, and with stepsize $\gamma := \frac{1}{L}$, GD yields

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2L}{T}\left(f(\mathbf{x}_0) - f(\mathbf{x}^\star)\right)$$

No overshooting If the function is smooth, and with the stepsize 1/L, gradient cannot overshoot (i.e. pass a critical point).

Trajectory analysis For non-convex functions, it's nice to have a good starting point and what happens when we start there.

Linear models with many outputs We have n data points, meaning n inputs $x_n \in \mathbb{R}^d$ and n outputs $y_n \in \mathbb{R}$. Hypothesis: $y_i \approx \mathbf{w}^\top x_i$ for some weight vector $\mathbf{w} \in \mathbb{R}^d$. We can generalize to more than one output value for each data point: n outputs $\mathbf{x}_n \in \mathbb{R}^m$. Hypothesis: $\mathbf{y}_i \approx W\mathbf{x}_i$ for a weight *matrix* $W \in \mathbb{R}^{m\times d}$.

Minimize LS To find that matrix $W^\star$ we pick the one that minimizes the least-squares error when taking into account all data points:

$$W^\star = \underset{W\in\mathbb{R}^{m\times d}}{\arg\min} \sum_{i=1}^{n} \|W\mathbf{x}_i - \mathbf{y}_i\|^2$$

Notation:

- $X \in \mathbb{R}^{d\times n}$: columns are the $\mathbf{x}_i$
- $Y \in \mathbb{R}^{m\times n}$: columns are the $\mathbf{y}_i$

This makes equivalent to compute

$$W^\star = \underset{W\in\mathbb{R}^{m\times d}}{\arg\min} \|WX - Y\|_F^2$$

This is cool for us: this argmin is a linear transformation $f(W)$. With that, the optimal is when the gradient is 0: $\nabla f(W^\star) = \mathbf{0}$. Equivalent to training a linear neural network with one layer under LS error.

Deep Linear NN Even with several layers (W_1, W_2, W_3), as long as the composition is linear, we can concatenate to $\mathbf{y} = W\mathbf{x}$, $W := W_3W_2w_1$.

Training with ℓ layers:

$$W^\star = \underset{W_1,S_2,\ldots,W_\ell}{\arg\min} \|W_\ell W_{\ell-1}\cdots W_1 X - Y\|_F^2$$

From here, we use a toy example: all matrices are 1×1, $W_i = x_i, X = 1, Y = 1, \ell = d$, so $f : \mathbb{R}^d \to \mathbb{R}$.

$$f(\mathbf{x}) := \frac{1}{2}\left(\prod_{k=1}^{d} x_k - 1\right)^2$$

Gradient With this toy f, the gradient is the following:

$$\nabla f(\mathbf{x} = \left(\prod_k x_k - 1\right)\left(\prod_{k\neq 1},\ldots,\prod_{k\neq d} x_k\right).$$

Balanced iterates Let $\mathbf{x} > \mathbf{0}$ (componentwise) and let $c \geq 1$ be a real number. $\mathbf{x}$ is called $c-$ *balanced* if $x_i \leq cx_j$ for all $1 \leq i, j \leq d$.

Lemma. *Let $\mathbf{x} \geq \mathbf{0}$ be c-balanced with $\prod_k x_k \leq 1$. Then for any stepsize $\gamma > 0$, $\mathbf{x}' := \mathbf{x} - \gamma\nabla f(\mathbf{x})$ satisfies $\mathbf{x}' \geq \mathbf{x}$ (componentwise) and is also c-balanced.*

Bounded Hessians

Lemma. *Suppose that* $\mathbf{x} > \mathbf{0}$ *is c-balanced. Then for any* $I \subseteq \{1, \dots, d\}$*, we have*

$$\begin{aligned} & \left(\tfrac{1}{c}\right)^{|I|} \left(\prod_k x_k\right)^{1-|I|/d} \\ \leq\ & \prod_{k \notin I} x_k \\ \leq\ & c^{|I|} \left(\prod_k x_k\right)^{1-|I|/d} \end{aligned}$$

A Tables and summaries

	GD	SGD
Lipschitz convex	$\mathcal{O}(\frac{1}{\epsilon^2})$	
Smooth & convex	$\mathcal{O}(\frac{1}{\epsilon})$	
Strongly convex		$\mathcal{O}(\frac{1}{\epsilon})$
Smooth & strongly convex	$\mathcal{O}(\log(\frac{1}{\epsilon}))$	
Subgradient		$\mathcal{O}(\frac{1}{\epsilon^2})$

Table 1: Convergence depending on function properties

B Useful functions

Euclidean Norm $\|X\|^2 = X^\top X = \sum_{n=1}^d x_i^2 \in \mathbb{R}_+$

Triangle inequality $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\|+\|\mathbf{y}\|$

Gradient

$$\nabla f(\mathbf{x}) := \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \dots, \frac{\partial f}{\partial x_d}(\mathbf{x})\right)$$

Hessian

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d}(\mathbf{x}) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_2 \partial x_2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d}(\mathbf{x}) \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1}(\mathbf{x}) & \frac{\partial^2 f}{\partial x_d \partial x_2}(\mathbf{x}) & \cdots & \frac{\partial^2 f}{\partial x_d \partial x_d}(\mathbf{x}) \end{pmatrix}$$

Positive semidefinite A symmetric matrix M is positive semidefinite if $\mathbf{x}^\top M \mathbf{x} \geq 0$ for all $\mathbf{x}$ and positive definite if $\mathbf{x}^\top M \mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$

Spectral norm Let A be an $(m \times d)$-matrix. Then

$$\|A\| := \max_{\mathbf{v} \in \mathbb{R}^d, \mathbf{v} \neq 0} \frac{\|A\mathbf{v}\|}{\|\mathbf{v}\|} = \max_{\|\mathbf{v}=1\|} \|A\mathbf{v}\|$$

is the 2-norm (or spectral norm) of A.

Frobenius Norm

$$\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$$

Equivalent to the euclidean norm of $vec(A)$, the "flattening" of A.