Optimization for Machine Learning 50.579

Instructor: Ioannis Panageas   Scribed by: Joel Weijia Lai

Lecture 2. Convex Optimization and Gradient Descent.

1 Introduction

In the previous week, we studied various gradient descent routine for differentiable functions. However, not all functions are differentiable at every point, as such, in order for our gradient descent routine to be applicable, we need to explore a new way of analyzing convex, non-differentiable functions.

1.1 Definitions

Definition 1.1 (Subgradients) *Let $f(x) : \mathcal{X} \to \mathbb{R}$ be a function, with $\mathcal{X} \subset \mathbb{R}^d$. $g_x \in \mathbb{R}^d$ is called a subgradient of f at x if for all $y \in \mathcal{X}$ we have*

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

Since the choice of subgradient is not unique, we can denote the set of subgradients at x by $\partial f(x)$ also called the subdifferential of f at the point x. Also, at a differentiable x, the only subgradient is $\nabla f(x)$.

Example 1.1 *Consider the function $f(x)$ given by Figure 1. $f(x)$ is clearly convex, but non-differentiable at x_0. Then, the subgradient at any x' that is differentiable is $\nabla f(x') = \frac{df(x')}{dx}$, while the subdifferential is given by the set of subgradients highlighted.*

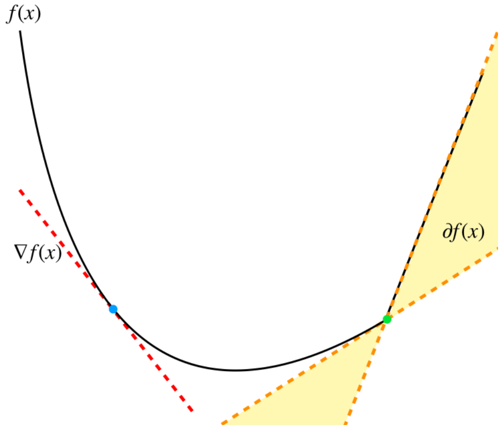

Figure 1: Figure for Example 1.1 of a convex function that is non-differentiable at a point.

Example 1.2 *Consider the function $f(x) = |x|$. $f(x)$ is clearly convex, but non-differentiable at $x = 0$. Then, the subgradient g_0 can be found by a direct application of the definition,*

$$\begin{aligned} |y| - |x| \geq& g_0(y - x), \\ |y| \geq& g_0 y. \end{aligned}$$

The subgradient g_0 satisfying the above inequality is $g_0 \in [-1, 1]$. To be precise,

$$\partial|x| = \begin{cases} 1 & x > 0, \\ -1 & x < 0, \\ [-1, 1] & x = 0. \end{cases}$$

Notice that in both examples, if $\mathbf{0} \in \partial f(x)$, then x is a global minimum. This is not a coincidence and will be the idea of Lemma 1.5, discussed later.

Theorem 1.3 (Supporting Hyperplane Theorem) *Let $C \subseteq \mathbb{R}^n$ be a* nonempty convex *set and $\bar{x}$ be a point on the boundary of C. Then, there exists a supporting hyperplane passing through $\bar{x}$ and containing the set C in one of its halfspaces.*

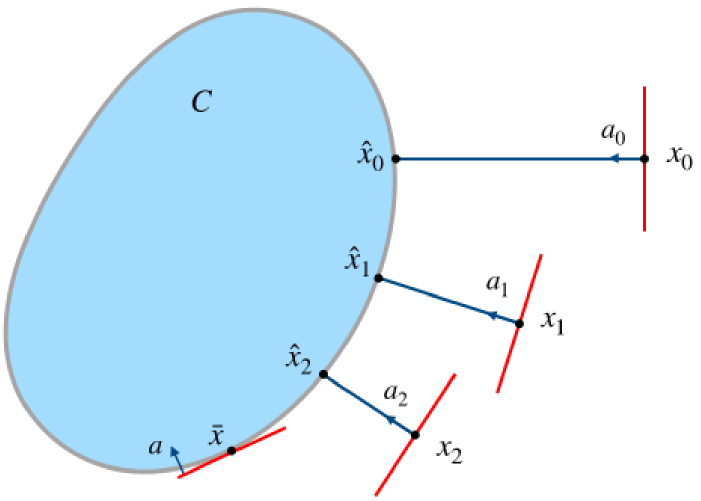

Figure 2: Proof by picture of the Supporting Hyperplane Theorem.

Proof: Figure 2 shows a pictorial proof of the theorem. Consider a sequence $\{x_k\}$ for $x_k \notin \mathrm{cl}(C)$ and converges to $\bar{x}$ as $k \to \infty$. Let $\hat{x}_k$ be the projection of x_k on $\mathrm{cl}(C)$. Then, we have,

$$a_k^\top x_k \leq a_k^\top x, \quad \forall x \in \mathrm{cl}(C), \forall k = 0, 1, \ldots.$$

where $a_k = (\hat{x}_k - x_k)/\|\hat{x}_k - x_k\|$. Let a be the limit point of $\{a_k\}$. ■

Lemma 1.4 (Existence and convexity) *Let $f : \mathcal{X} \to \mathbb{R}$ be a function such that $\partial f(x) \neq \varnothing$ for all x. Then, it holds that f is convex.*

Proof: For the choice: $x := ty + (1-t)x$ and $y := x$, then it holds that there exists a vector g such that

$$f(ty + (1-t)x) - f(x) \leq g^\top t(y - x). \tag{1}$$

In the same way, for the choice: $y := ty + (1-t)x$ and $x := y$, it holds that there exists a vector g such that

$$f(ty + (1-t)x) - f(y) \leq g^\top (1-t)(x - y). \tag{2}$$

Note that the inequalities are reversed. Then if we were to take the linear combination $(1-t) \cdot (1) + t \cdot (2)$, we have

$$f(ty + (1-t)x) \leq tf(y) + (1-t)f(x).$$

This is the definition of convexity that we have seen in Lecture 1. The converse is also true. This is a consequence of the Supporting Hyperplane Theorem. ■

Lemma 1.5 (Local minima are global minima) *Let $f : \mathcal{X} \to \mathbb{R}$ be a convex function. If x is a local minimum, then it is a global minimum. This happens if and only if $\mathbf{0} \in \partial f(x)$.*

This lemma claims the uniqueness of the minimum. The proof is for x being a global minimum if and only if $\mathbf{0} \in \partial f(x)$ is straight forward. We shall prove the uniqueness of the minimum.
Proof: For a small enough $t > 0$. Suppose there are local minima at x and y. Then

$$\begin{aligned} f(x) &\leq f(tx + (1-t)y) \\ &\leq tf(x) + (1-t)f(y), \\ \Rightarrow (1-t)f(x) &\leq (1-t)f(y), \end{aligned}$$

hence, we conclude that $f(x) \leq f(y)$. However, x and y are arbitrary, so the same can be said for $f(y) \leq f(x)$. Thus, we conclude that $f(x) = f(y)$, the minimum is unique. ■

2 Gradient Descent Revisited

Definition 2.1 (Gradient Descent (Subgradient)) *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function that is not necessarily differentiable in some convex set $\mathcal{X}$. The algorithm is defined iteratively:*

$$x_{k+1} = x_k - \alpha g_{x_k},$$

where $g_{x_k} \in \partial f(x_k)$ is the subgradient computed at x_k.

Note that this definition is a slight modification from the definition we saw in Lecture 1. (1) The function is not necessarily differentiable in the convex set $\mathcal{X}$. (2) The gradient $\nabla f(x_k)$ is replaced by the subgradient. However, we have the same guarantees as discussed in classic and projected gradient descent. Here, we discuss the analysis of gradient descent for L-Lipschitz.

Exercise 3 (General case). *Suppose $f(x)$ is L-Lipschitz continuous and $\partial f(x) \neq \varnothing$. Then $\forall x \in dom(f)$*

$$\|g_x\|_2 \leq L \textit{ where } g_x \in \partial f(x).$$

This the the general case of Exercise 3 from Lecture 1. Prove is left as an exercise.

Theorem 2.1 (Gradient Descent (Subgradient)) *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex (want to minimize) and L-Lipschitz. Let $R = \|x_1 - x^*\|_2$, the distance between the initial point x_1 and minimizer x^*. It holds for $T = \frac{R^2L^2}{\epsilon^2}$,*

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \epsilon,$$

with approximately choosing $\alpha = \frac{\epsilon}{L^2}$

Proof:

$$\begin{aligned}
f(x_t) - f(x^*) &\leq g_{x_t}^\top (x_t - x^*) && \text{(def. subgradient)}\\
&= \frac{1}{\alpha}(x_t - x_{t+1})^\top (x_t - x^*) && \text{(def. gradient descent 2.1)}\\
&= \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 + \|x_t - x_{t+1}\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) && (2a^\top b = |a|^2 + |b|^2 - |a-b|^2)\\
&= \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\alpha}{2}\|g_{x_t}\|_2^2 && \text{(def. gradient descent 2.1)}\\
&\leq \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\alpha L^2}{2}. && \text{(Exercise 3)}
\end{aligned}$$

Now, we take the telescopic sum across $t = 1, \ldots, T$ and dividing by T. We get

$$\frac{1}{T}\sum_{t=1}^{T}\left(f(x_t) - f(x^*)\right) \leq \frac{1}{2\alpha T}\sum_{t=1}^{T}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\alpha L^2}{2}.$$

For the LHS, we apply Jensen's inequality

$$\frac{1}{T}\sum_{t=1}^{T}\left(f(x_t) - f(x^*)\right) \geq f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*).$$

For the RHS, taking the telescopic sum, we have

$$\begin{aligned}
\frac{1}{2\alpha T}\sum_{t=1}^{T}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) &\leq \frac{1}{2\alpha T}\left(\|x_1 - x^*\|_2^2 - \|x_{T+1} - x^*\|_2^2\right)\\
&\leq \frac{1}{2\alpha T}\|x_1 - x^*\|_2^2,
\end{aligned}$$

as the second term is non-negative. Thus, we put every thing together

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \frac{1}{2\alpha T}\|x_1 - x^*\|_2^2 + \frac{\alpha L^2}{2} = \frac{R^2}{2\alpha T} + \frac{\alpha L^2}{2} = \epsilon,$$

for the choice $\alpha = \dfrac{\epsilon}{L^2}$ and $T = \dfrac{R^2L^2}{\epsilon^2}$. ■

3 Stochastic Gradient Descent (SGD)

Definition 3.1 (Stochastic Gradient Descent) *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex (want to minimize). The SGD algorithm is defined iteratively:*

$$x_{k+1} = x_k - \alpha_k v_k,$$

where $\mathbb{E}[v_k|x_k] \in \partial f(x_k)$.

Key remarks:

- α_k is called the stepsize. Intuitively, the smaller, the slower the algorithm.
- α_k must depend on k, i.e $\alpha_k \to 0$ as $k \to \infty$.
- v_k and x_k are random variables. In SGD v_k is updated randomly, i.e $v_k = \nabla f(x_k) + \zeta_k$ for $\mathbb{E}[\zeta_k|x_k] = 0$

Theorem 3.1 (μ-Strongly Convex Stochastic Gradient Descent) *Let $f : \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex (want to minimize). Moreover, assume that $\mathbb{E}[\|v_k\|^2] \leq \rho^2$. Let x^* be the minimizer. It holds for $\alpha_k = \frac{1}{\mu k}$,*

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_t x_t\right)\right] - f(x^*) \leq \frac{\rho^2}{2\mu T}(1 + \log T).$$

Key remarks:

- α_k scales as $\frac{1}{k}$ and is vanishing to talk about convegence.
- For $T = \Theta\left(\frac{1}{\epsilon}\log\frac{1}{\epsilon}\right)$ we get an error ϵ.
- Rakhlin, Shamir & Sridharan (2012) derived a convergence rate in which $\log T$ is eliminated for a variant [1].
- Shamir & Zhang (2013) showed theorem above for last iterate for last iterate x_T, i.e $\mathbb{E}[f(x_t)] - f(x^*) \leq \frac{\rho^2}{2\mu T}(1 + \log T)$ [2].

Proof: Set $\nabla^t = \mathbb{E}[v_t|x_t]$ (this is a random variable). From strong convexity, we get

$$(x_t - x^*)^\top \nabla^t \geq f(x_t) - f(x^*) + \frac{\mu}{2}\|x_t - x^*\|_2^2,$$
$$\mathbb{E}[(x_t - x^*)^\top \nabla^t] \geq \mathbb{E}\Big[f(x_t) - f(x^*) + \frac{\mu}{2}\|x_t - x^*\|_2^2\Big].$$

This will form the lower bound. For the upper bound, we claim

$$\mathbb{E}[(x_t - x^*)^\top \nabla^t] \leq \frac{\mathbb{E}[\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2]}{2\alpha_t} + \frac{\alpha_t \rho^2}{2}.$$

Proof of claim: The Law of Cosines gives

$$\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2 \geq 2\alpha_t (x_t - x^*)^\top v_t - a_t^2 \|v_t\|_2^2.$$

Taking the expectation on both sides,

$$\mathbb{E}\big[\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\big] \geq \mathbb{E}\big[2\alpha_t (x_t - x^*)^\top v_t\big] - \mathbb{E}\big[a_t^2 \|v_t\|_2^2\big],$$

and, for the first term of the RHS,

$$\begin{aligned} \mathbb{E}\big[\mathbb{E}[2\alpha_t (x_t - x^*)^\top v_t | x_t]\big] &= \mathbb{E}\big[2\alpha_t (x_t - x^*)^\top \mathbb{E}[v_t | x_t]\big] \\ &= \mathbb{E}\big[2\alpha_t (x_t - x^*)^\top \nabla^t\big]. \end{aligned}$$

A rearrangement of the terms will give the inequality of the claim. ■

Thus, if we were to put the lower and upper bound together and apply linear expectation, i.e $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$

$$\begin{aligned} \mathbb{E}\Big[f(x_t) - f(x^*) + \frac{\mu}{2}\|x_t - x^*\|_2^2\Big] &\leq \frac{\mathbb{E}[\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2]}{2\alpha_t} + \frac{\alpha_t \rho^2}{2}, \\ \mathbb{E}[f(x_t) - f(x^*)] &\leq \frac{\mathbb{E}[\|x_t - x^*\|_2^2 (1 - \alpha_t \mu) - \|x_{t+1} - x^*\|_2^2]}{2\alpha_t} + \frac{\alpha_t \rho^2}{2}. \end{aligned}$$

Therefore, we take the telescopic sum across t and dividing by T, and recall that $\alpha_t = \frac{1}{t\mu}$,

$$\begin{aligned} \mathbb{E}\left[\frac{1}{T}\sum_t f(x_t)\right] - f(x^*) &\leq \mathbb{E}\left[-\frac{\mu T}{2}\|x_T - x^*\|_2^2\right] + \frac{\rho^2}{2\mu}\frac{1}{T}\sum_t \frac{1}{t} \\ &\leq \frac{\rho^2}{2\mu}\frac{1}{T}\sum_t \frac{1}{t}, \end{aligned}$$

since the first term on the RHS is non-positive. Applying Jensen's inequality, and the fact that $\sum_{t=1}^T \frac{1}{t} \leq 1 + \log T$, we have the result for μ-strongly convex SGD

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_t x_t\right)\right] - f(x^*) \leq \frac{\rho^2}{2\mu T}(1 + \log T).$$

■

Theorem 3.2 (General Stochastic Gradient Descent) *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function (want to minimize). Moreover, assume that $\|v_k\|^2 \leq \rho$ with probability one. Let $R = \|x_1 - x^*\|_2$, the distance between the initial point x_1 and minimizer x^*. It holds for $\alpha = \frac{R}{\rho\sqrt{T}}$,*

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_t x_t\right)\right] - f(x^*) \leq \frac{R\rho}{\sqrt{T}}.$$

Key remarks:

- α is fixed but scales as $\sqrt{\frac{1}{T}}$ is vanishing to talk about convergence.
- For $T = \Theta\left(\frac{1}{\epsilon^2}\right)$ we get an error ϵ.

Proof: As a notation, we denote $\mathbb{E}_{1:k}[\cdot]$ as the expectation of the joint distribution of random variables $(v_1, \ldots, v_k)$.

$$\begin{aligned}
\mathbb{E}_{1:T}[f(x_t) - f(x^*)] &\leq \mathbb{E}_{1:T}[(x_t - x^*)^\top \nabla^t] \\
&= \mathbb{E}_{1:t-1}\left[\mathbb{E}_{1:T}[(x_t - x^*)^\top \nabla^t | v_1, \ldots, v_{t-1}]\right] && \text{(Conditional expectation)} \\
&= \mathbb{E}_{1:T}\left[(x_t - x^*)^\top \mathbb{E}_{1:t-1}[\nabla^t | v_1, \ldots, v_{t-1}]\right] && \text{(Deterministic in } v_1, \ldots . v_{t-1}\text{)} \\
&= \mathbb{E}_{1:T}\left[(x_t - x^*)^\top v_t\right] \\
&\leq \mathbb{E}_{1:T}\left[\frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 + \|x_t - x_{t+1}\|_2^2 - \|x_{t+1} - x^*\|_2^2\right)\right] && (2a^\top b = |a|^2 + |b|^2 - |a-b|^2) \\
&= \mathbb{E}_{1:T}\left[\frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right)\right] + \frac{\alpha\|v_t\|_2^2}{2} && \text{(def. of SGD 3.1)} \\
&\leq \mathbb{E}_{1:T}\left[\frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right)\right] + \frac{\alpha\rho 2}{2}. && (\|v_t\|_2 \leq \rho)
\end{aligned}$$

Therefore, we take the telescopic sum across t and dividing by T,

$$\begin{aligned}
\mathbb{E}_{1:T}\left[\frac{1}{T}\sum_t f(x_t)\right] - f(x^*) &\leq \frac{1}{2\alpha T}\mathbb{E}_{1:T}[\|x_1 - x^*\|_2^2 - \|x_{T+1} - x^*\|_2^2] + \frac{\alpha\rho^2}{2} \\
&\leq \frac{R^2}{2\alpha T} + \frac{\alpha\rho^2}{2} = \frac{R\rho}{\sqrt{T}}
\end{aligned}$$

again, we apply Jensen's inequality and noting that $\alpha = \frac{R}{\rho\sqrt{T}}$, we arrive at the result for the general SGD

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_t x_t\right)\right] - f(x^*) \leq \frac{R\rho}{\sqrt{T}}.$$

■

Definition 3.2 (Coordinate Descent) *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex differentiable function in some convex set $\mathcal{X}$. The CD algorithm is defined iteratively by choosing a coordinate i, where i is drawn uniformly at random from $[d]$ and update:*

$$x_{k+1} = x_k - \alpha_k \frac{\partial f(x_k)}{\partial x_i} \cdot e_i,$$

where e_i is a unit vector with 1 in position i and zeros in all other positions.

Once can view Coordinate Descent as a Stochastic Gradient Descent with the specific oracle $\tilde{g}(x) = d\frac{\partial f(x)}{\partial x_j} \cdot e_j$, where j is drawn uniformly at random from $[d]$. Clearly, $\mathbb{E}[\tilde{g}(x)] = \nabla f(x)$, and furthermore $\mathbb{E}\left[\|\tilde{g}(x)\|_2^2\right] = d\|\nabla f(x)\|_2^2$.

4 Stochastic Gradient Descent (Examples)

Having seen how SGD works, it is not imperative that we see the kinds of problems that SGD can solve. One such instance the use of SGD to solve risk minimization problems, also called Maximum Likelihood Estimates (MLE) problems.

4.1 Risk Minimization

Definition 4.1 (Risk Minimization) *Let $l(x, z) : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ be a risk function and D some unknown distribution we can get samples from. We are interested in solving:*

$$\min_{x \in \mathcal{X}} L(x), \ \textit{where } L(x) := \mathbb{E}_{z \sim D}[l(x, z)].$$

There are two approaches to tackling such problems,

1. Take a large number (say n) samples z_i independently and consider the estimate $\bar{L}(x) := \frac{1}{n} \sum_i l(x, z)$. By the Law of Large Numbers this is a close enough (hopefully) estimate of $L(x)$. Then, we run a first order optimization algorithm (say GD) on $\bar{L}(x)$. This is a possible means of solving the problem. However, if we do not know the form of $l(x, z)$, then we are essentially stuck. Also, this requires many calculations (n) to perform one optimization step.

2. Of course, the second method is to perform Stochastic Gradient Descent!. For each iteration $t+1$, take a new sample z_t independently from $z_1, \ldots, z_{t-1}$ and consider the unbiased estimate $\nabla_x l(x, z)$. Then, we update $x_{t+1} = x_t - \alpha_t \nabla_x l(x, z)$.

4.2 Examples

Example 4.1 (MLE for Gaussian) *Let $z \sim \mathcal{N}(\mu, 1)$ and $l(x, z) := -\log p_x(z)$ denotes the log-likelihood of $mathcalN(x, 1)$. Here, we have an unknown distribution D, which is a Gaussian with unknown mean, μ. We are interested in solving:*

$$\min_{x \in \mathbb{R}} \mathbb{E}_{z \sim \mathcal{N}(\mu,1)}[-\log p_x(z)],$$

where $p_x(z)$ is the probability density function.

Some remarks on Maximum (log)-Likelihood:

1. The standard approach for parameter estimation boils down to creating an optimization problem that best solves for parametric families of distributions.

2. Under certain assumptions, the Maximum (log) Likelihood estimator is consistent! The minimizer should reveal some information about the "unknowns".

3. Since the probability density function $p_x(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(z-x)^2}{2}\right]$, the above problem boils down to $\min_{x \in \mathbb{R}} \mathbb{E}_{z \sim \mathcal{N}(\mu,1)} \left[\frac{(z-x)^2}{2}\right]$.

It is clear that the minimizer $x^* = \mu$. However, we do not know what is μ, so we start by solving the simplified minimization problem:

$$\min_{x\in\mathbb{R}} \mathbb{E}_{z\sim\mathcal{N}(\mu,1)}\Big[\frac{(z-x)^2}{2}\Big].$$

- The derivative is $(x-z)$, i.e. $v_t = (x_t - z_t)$
- $\mathbb{E}_z\big[(x-z)^2\big] = 1 + (x-\mu)^2$. This is bounded, which is exactly what we require from SGD, $\mathbb{E}\big[\|v\|_2^2\big] \le \rho^2$.
- The second derivative is 1, hence, this is 1-strongly convex.
- Starting from $x_0 = 0$, at iteration $t+1$, we get a fresh sample z_t and we have $x_{t+1} = x_t - \alpha_t(x_t - z_t)$

Choosing $\alpha_t = \frac{1}{t}$, we can check that $x_T = \frac{1}{t}\sum_{t=1}^T z_t$. This is the empirical mean.

Question 1: We know that $f(\tilde{x}) - f(x) \le \epsilon$. But what is $\|\tilde{x} - x\|$ bounded by? **Answer 1:** Consider strong convexity:

$$\epsilon' \ge f(\tilde{x}) - f(x) \ge \nabla f(\mu)(\tilde{x} - \mu) + \frac{1}{2}\|\tilde{x} - \mu\|_2^2$$

Here, $\tilde{x}$ is a sample drawn from the distribution of the empirical mean, i.e.

$$\frac{1}{T}\sum_{t=1}^{T} z_t \sim \mathcal{N}\Big(\mu, \frac{1}{T}\Big).$$

Thus, $\epsilon' \ge \|\tilde{x} - \mu\|_2^2$. The first term vanishes because $\nabla f(\mu) = 0$. Thus, $\|\tilde{x} - \mu\|_2 \le \sqrt{\epsilon'} \equiv \epsilon \Rightarrow \|\tilde{x} - \mu\|_2 \le \epsilon^2$. Recall, for $T = \Theta\Big(\frac{1}{\epsilon'}\log\frac{1}{\epsilon'}\Big)$, we get an error of ϵ. Hence, we get $T \sim \frac{1}{\epsilon^2}\log\frac{1}{\epsilon^2}$.

Question 2: Having shown that can get ϵ-close to μ after $\frac{1}{\epsilon^2}\log\frac{1}{\epsilon^2}$ iterations, we claim that this is not the best we can do. Why? **Answer 2:** To see why, we know that for a Gaussian distribution, the probability of a sample falling beyond 3 standard deviations from the mean is about 1%, i.e. $\mathbb{P}\Big(|\tilde{x} - \mu| \ge \frac{3}{\sqrt{T}} \equiv \epsilon\Big) \sim 1\%$. Therefore, we conclude that $T \sim \frac{1}{\epsilon^2}$. Note that it might not always be the case that the empirical mean can give us information about the distribution $\mathcal{D}$.

Example 4.2 (Bias of a coin) *Assume you are given a coin that gives H with probability $p \in (0,1)$ and T with probability $1-p$. How many tosses do you need to get an estimate $\tilde{p}$ about p and be sure with probability 99% that $|p - \tilde{p}| \le \epsilon$? [Hint: $f_p(z) = p^z(1-p)^{1-z}$]*

It is clear that the minimizer $x^* = p$. However, we do not know what is p, so we start by solving the minimization problem:

$$\min_{x} \mathbb{E}\big[-z\log x - (1-z)\log(1-x)\big].$$

- The derivative of l is $-\frac{z}{x}+\frac{1-z}{1-x}=\frac{x-z}{x(1-x)}$, which has absolute value at most $\frac{1}{\epsilon}$ for $x\in(\epsilon,1-\epsilon)$.
- The second derivative of L is $\frac{p}{x^2}+\frac{1-p}{(1-x)^2}$, hence, it is $4(p-p^2)$-strongly convex for $x\in(0,1)$. Notice that $(p-p^2)$ is the variance of the Bernoulli distribution.
- Starting from $x_0=1/2$, at iteration $t+1$, we get a fresh sample z_t and we have $x_{t+1}=x_t-\alpha_t\frac{x_t-z_t}{x_t(1-x_t)}$.

As seen in the previous example, $T\sim\frac{1}{\epsilon}\log\frac{1}{\epsilon}$, for μ-strongly convex SGD, is asymptotically equivalent to $\frac{\rho^2}{2\mu\epsilon}\log\frac{1}{\epsilon}$. We have seen that $\rho=\frac{1}{\epsilon}$ and $\mu=4(p-p^2)$. Thus, we conclude that we can get ϵ'-close to the log-likelihood after $\frac{1}{(p-p^2)\epsilon'^3}\log\frac{1}{\epsilon'}$ iterations and ϵ-close to p after $\frac{1}{(p-p^2)\epsilon^6}\log\frac{1}{\epsilon^2}$ iterations. We ask ourselves the same question: Can we do better? Again, yes. Since $z_t\sim\mathcal{B}(p)$, we can calculate the mean and variance of the empirical mean,

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{T}\sum z_t\Big]&=\frac{1}{T}\sum\mathbb{E}[z_t]\\&=\frac{1}{T}\sum 1\cdot p+0(1+p)\\&=p\end{aligned}$$

$$\begin{aligned}\mathrm{Var}\Big[\sum z_t\Big]&=\sum\mathrm{Var}[z_t]\\&=\sum p-p^2=T(p-p^2)\\\Rightarrow\mathrm{Var}\Big[\frac{1}{T}\sum z_t\Big]&=\frac{1}{T^2}\mathrm{Var}\Big[\sum z_t\Big]=\frac{1}{T}(p-p^2)\end{aligned}$$

Thus, having obtained the variance, to have a 99% confidence, we apply the Chebyshev's inequality,

$$\begin{aligned}\mathbb{P}\Big[\Big|\frac{1}{T}\sum z_t\Big|^2\geq\epsilon^2\Big]\leq\frac{\mathrm{Var}\Big[\frac{1}{T}\sum z_t\Big]}{\epsilon^2}&=1\%\\\Rightarrow\frac{p-p^2}{T}&=\frac{\epsilon^2}{100}\\\Rightarrow T\sim O\Big(\frac{p-p^2}{\epsilon^2}\Big)&\sim O\Big(\frac{1}{\epsilon^2}\Big)\end{aligned}$$

Example 4.3 (Non-example: Mixture of Gaussians) *Assume you have access to i.i.d samples from $z\sim\mathcal{N}(\mu,1)$. However, there is an adversary that with probability 1/2 corrupts z and gives you $-z$. Can you infer/estimate μ?*

To answer the question, we need to solve:

$$\min_{x\in\mathbb{R}}\mathbb{E}_{z\sim\mathcal{N}(\mu,1)}\Bigg[-\log\Big(\frac{1}{2\sqrt{2\pi}}\exp\Big[-\frac{(z-x)^2}{2}\Big]+\frac{1}{2\sqrt{2\pi}}\exp\Big[-\frac{(z+x)^2}{2}\Big]\Big)\Bigg].$$

This, however is not convex, the proof is left as an exercise (Hint: suffice to show that the gradient at $x=0$ is 0). While the empirical mean is zero, it does not reveal any information about μ. Stochastic Gradient Descent cannot help us here. This is the motivation for the content covered in subsequent weeks where we start to look at non-convex functions.

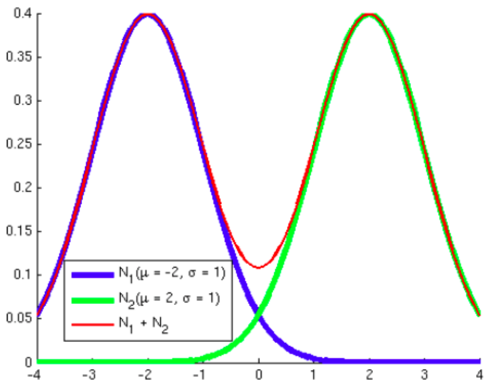

Figure 3: Probability density function of the mixture of 2 Gaussians

References

[1] Rakhlin, A., Shamir, O., & Sridharan, K. (2011). *Making gradient descent optimal for strongly convex stochastic optimization.* arXiv preprint arXiv:1109.5647.

[2] Shamir, O., & Zhang, T. (2013). *Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes.* International conference on machine learning.