Chapter 7: Optimization for learning

1. Two paradigms: empirical risk minimization vs stochastic approximation
2. Illustration: gradient descent for least squares
3. Convexity: an interlude
4. Gradient descent
5. Accelerated gradient descent
6. Projected subgradient method
7. Minimax lower bound for deterministic convex optimization
8. Stochastic gradient method and Polyak-Juditsky averaging
9. Stochastic variance reduced gradient (SVRG)

Two paradigms: Empirical Risk Minimization & Stochastic Approximation

We have seen that many learning tasks, such as in regression and maximum likelihood estimation, amount to a stochastic optimization problem:

$$\min_{x \in \mathbb{R}^d} \ \mathop{\mathbb{E}}_{z \sim \mathcal{P}} \ell(x, z). \tag{1}$$

In this chapter, we will discuss algorithms for solving such problems.

Two paradigms: Empirical Risk Minimization & Stochastic Approximation

There are essentially two strategies, which yield similar guarantees.

Strategy 1 (Empirical Risk Minimization): Draw $z_1, \ldots, z_n \overset{iid}{\sim} \mathcal{P}$ and declare

$$x_n = \arg\min_x \frac{1}{n} \sum_{i=1}^{n} \ell(x, z).$$

There are variants where one would add a regularizer (e.g. ridge regression) or impose a constraint on x. A key observation is that when forming the ERM, an error on the order of $1/n$ or $1/\sqrt{n}$ is already incurred for the true problem (1) to be solved. Therefore one should not solve ERM to higher accuracy than this "estimation error", lest one "overfits" to the observed data.

Strategy 2 (Stochastic approximation): These are algorithms that proceed in each iteration t by drawing a single sample $z_t \sim \mathcal{P}$ and taking a step from x_t using some information gathered from the random function $f(\cdot, z_t)$. A prime example is the stochastic gradient method, which we will discuss in detail.

Gradient descent for least squares

As a warm, consider the least squares objective

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{2n} \|Ax - b\|_2^2.$$

Let $\bar{x}$ be a minimizer of f and set $f^* = f(\bar{x})$. Optimality conditions imply

$$A^\top A \bar{x} = A^\top b$$

The simple gradient descent algorithm takes the form

$$x_{t+1} = x_t - \frac{\eta}{n} A^\top (A x_t - b),$$

for some parameter $\eta > 0$ to be chosen. Note that we may equivalently write

$$x_{t+1} - \bar{x} = \left(I - \frac{\eta}{n} A^\top A\right)(x_t - \bar{x}).$$

Gradient descent for least squares

Setting $H = \frac{1}{n} A^\top A$ we further arrive at

$$\boxed{x_t - \bar{x} = (I - \eta H)^t \, (x_0 - \bar{x})} \tag{2}$$

Since f is a pure quadratic, we may write

$$f(x) = f(\bar{x}) + \langle \underbrace{\nabla f(\bar{x})}_{=0}, x - \bar{x} \rangle + \frac{1}{2} \langle \underbrace{\nabla^2 f(\bar{x})}_{=H} (x - \bar{x}), x - \bar{x} \rangle.$$

Thus, we conclude

$$\boxed{f(x_t) - f^* = \frac{1}{2} (x_0 - \bar{x})^\top \, (I - \eta H)^{2t} \, H (x_0 - \bar{x})} \tag{3}$$

Set $\beta = \lambda_{\max}(H)$, $\alpha = \lambda_{\min}(H)$, and define the condition number $\kappa = \frac{\beta}{\alpha}$.

Gradient descent for least squares

Let us analyze the decay of (2) and (3), beginning with the former:

$$\|x_t - \bar{x}\|_2^2 \le \left(\max_{\lambda \in [\alpha,\beta]} |1 - \eta\lambda|\right)^{2t} \|x_0 - \bar{x}\|^2.$$

It is easy to see that $\min_{\eta>0} \max_{\lambda\in[\alpha,\beta]} |1-\eta\lambda|$ is attained by $\eta = \frac{2}{\alpha+\beta}$ thereby yielding the linear rate $\frac{\kappa-1}{\kappa+1}$. Since $\alpha > 0$ is often difficult to estimate, it suffices to choose $\eta = \frac{1}{\beta}$ which results in the same rate up to a constant:

$$\boxed{\|x_t - \bar{x}\|_2^2 \le \left(1 - \kappa^{-1}\right)^{2t} \|x_0 - \bar{x}\|^2}.$$

We may further upper bound the right side by $\exp(-t/\kappa)\|x_0 - \bar{x}\|^2$. Setting this quantity to ϵ, we see that it suffices to perform $t = \kappa \cdot \log(\|x_0 - \bar{x}\|^2/\epsilon)$ iterations to find a point x satisfying $\|x - \bar{x}\|^2 \le \epsilon$.

A similar argument with the step-size $\eta = \frac{1}{\beta}$ shows

$$\boxed{f(x_t) - f^* \le \left(1 - \frac{1}{\kappa}\right)^{2t} (f(x_0) - f^*)}.$$

Gradient descent for least squares

The convergence rates we have obtained are highly sensitive to κ, and in particular to α, which is typically on the order of n^{-1} or $n^{-1/2}$. Let us next show how to obtain a rate that is insensitive to α, but which is sublinear in t.

From (3) we have

$$f(x_t) - f^* \le \frac{1}{2} \max_{\lambda \in [\alpha, \beta]} |\lambda(1 - \lambda/\beta)^{2t}| \cdot \|x_0 - \bar{x}\|^2$$

Observe $|\lambda(1-\lambda/\beta)^t| \le \lambda \exp(-\lambda/\beta)^{2t} = \frac{\beta}{2t}\frac{2t\lambda}{\beta}\exp(-2t\lambda/\beta) \le \frac{\beta}{2te}$ where we used that $\max_{s \ge 0} se^{-s} = e^{-1}$. Thus we conclude

$$\boxed{f(x_t) - f^* \le \frac{\beta\|x_0 - \bar{x}\|^2}{8t}}.$$

Our next goal is to develop similar guarantees for gradient type methods beyond least squares.

Gradient descent for smooth minimization

Will aim to minimize a C^1-smooth function f on $\mathbb{R}^d$ by the gradient method:

$$\boxed{x_{t+1} = x_t - \eta \nabla f(x_t)}$$

where $\eta > 0$ is to be chosen. Suppose f has a minimizer $\bar{x}$ and set $f^\star := f(\bar{x})$.

In order to make progress it will be important to quantify "how smooth" is f.

Definition (Quantifying smoothness)

A function $f\colon \mathbb{R}^d \to \mathbb{R}$ is called β-smooth if it is C^1-smooth and satisfies

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \cdot \|x - y\| \qquad \forall x, y.$$

You will check the following for homework.

Lemma: A C^2-smooth function f is β-smooth if and only if $\nabla^2 f(x) \preceq \beta \cdot I_d$ for all $x \in \mathbb{R}^d$.

Gradient descent for smooth minimization

In order to analyze gradient descent, we will need the following.

Corollary (Accuracy in approximation)

Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ *is a* β*-smooth function. Then for any points* $x, y \in \mathbb{R}^d$ *the inequality*

$$\left| f(y) - f(x) - \langle \nabla f(x), y - x \rangle \right| \leq \frac{\beta}{2} \|y - x\|^2 \quad \textit{holds.} \tag{4}$$

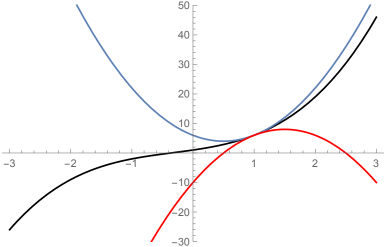


Figure: The black curve depicts the graph of a β-smooth function f; the blue and red curves depict graphs of the quadratics $Q_1(y) = f(x) + \nabla f(x), y - x\rangle + \frac{\beta}{2}\|y - x\|^2$ and $Q_2(y) = f(x) + \nabla f(x), y - x\rangle - \frac{\beta}{2}\|y - x\|^2$, respectively.

Proof

Fix $x, y \in \mathbb{R}^d$ and define the function $\varphi(t) = f(x + t(y - x))$. Then the fundamental theorem of calculus gives

$$\begin{aligned} \varphi(1) &= \varphi(0) + \int_0^1 \varphi'(t)\,dt \\ &= \varphi(0) + \varphi'(0) + \int_0^1 (\varphi'(t) - \varphi'(0))\,dt. \end{aligned}$$

Noting the equality $\varphi'(t) = \langle \nabla f(x + t(y - x)), y - x\rangle$, we deduce $|\varphi'(t) - \varphi'(0)| \leq \beta \|y - x\|^2 \cdot t$, thereby completing the proof. □

Gradient descent for smooth minimization

Setting $y = x - \eta \nabla f(x)$ yields an estimate on functional improvement.

Lemma (Descent)

The gradient step $x^+ = x - \eta \nabla f(x)$ *satisfies*

$$f(x) - f(x^+) \geq \eta \left(1 - \frac{\eta\beta}{2}\right) \|\nabla f(x)\|^2.$$

The term $\eta \left(1 - \frac{\eta\beta}{2}\right)$ is maximized by setting $\eta = \frac{1}{\beta}$, yielding

$$\boxed{f(x) - f(x^+) \geq \frac{1}{2\beta} \|\nabla f(x)\|^2}.$$

Theorem (Complexity)

Suppose f *is* β*-smooth. Then gradient descent iterates* x_t *with* $\eta = \frac{1}{\beta}$ *satisfy*

$$\min_{i=1,\ldots,t} \|\nabla f(x_i)\|^2 \leq \frac{1}{t} \sum_{i=1}^{t} \|\nabla f(x_i)\|^2 \leq \frac{2\beta(f(x_1) - f^\star)}{t}$$

Proof

From the descent lemma, we have

$$f(x_1) - f^\star \geq f(x_1) - f(x_{t+1}) = \sum_{i=1}^{t} f(x_i) - f(x_{i+1}) \geq \frac{1}{2\beta} \sum_{i=1}^{t} \|\nabla f(x_i)\|^2.$$

Dividing both sides by t and using that the average of t positive numbers is bigger than their minimum completes the proof. □

Convexity

Gradient descent turns to be much faster for convex problems.

Definition (Convexity)

A function $f: \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is called convex if it satisfies the secant inequality

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \qquad \forall x, y \in \mathbb{R}^d, \lambda \in [0,1].$$

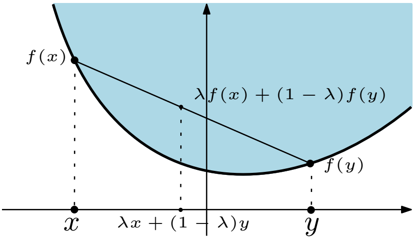


Figure: Secant inequality.

More generally, we say that f is α-strongly convex if the perturbed function $g(x) = f(x) - \frac{\alpha}{2}\|x\|^2$ is convex.

Preservation of convexity

Convexity is preserved under the following operations (check this!).

1. If f is convex and $\lambda \geq 0$, then $g(x) = \lambda f(x)$ is convex.
2. If f and g are convex, the $f + g$ is convex
3. If f is convex, then $g(y) = f(Ay)$ is convex for any linear map A.
4. If f_i are convex for all $i \in \mathcal{I}$, where $\mathcal{I}$ is an arbitrary set, then the function $f(x) = \sup_{i \in \mathcal{I}} f_i(x)$ is convex.
5. If $f(x, y)$ is convex, then so is the function $g(x) = \inf_y \; f(x, y)$.
6. If f is convex and A is a linear map, then the following function is convex:

$$g(x) = \inf_y \; \{f(y) : \text{subject to } Ay = x\}.$$

Convexity and tangent lines

We will need the following characterization of smooth convex functions in terms of tangent lines.

Theorem (Convexity and tangent lines)

A C^1-smooth function f is α-strongly convex if and only if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|x - y\|^2 \qquad \forall x, y. \tag{5}$$

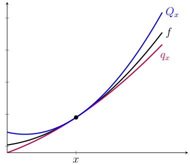


Figure: $Q_x(y) := f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2}\|y - x\|^2$ is an upper estimator based at x and $q_x(y) := f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|^2$ is a lower estimator based at x.

In particular, if $\bar{x}$ is a minimizer of a α-strongly convex function g, then

$$g(x) \geq g(\bar{x}) + \frac{\alpha}{2}\|x - \bar{x}\|^2.$$

We will use this often in convergence proofs with certain auxiliary functions g!

Proof

It suffices to establish the theorem with $\mu = 0$, since the general statement follows by applying it to $f - \frac{\mu}{2}\|\cdot\|^2$. Suppose first that f is convex. Then for any $t \in (0,1)$, convexity implies

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

while the definition of the derivative yields

$$f(x + t(y-x)) = f(x) + t\langle \nabla f(x), y - x\rangle + o(t).$$

Combining the two expressions and dividing by t yields the relation

$$f(y) - f(x) \geq \langle \nabla f(x), y - x\rangle + o(t)/t.$$

Letting t tend to zero yields property (5). Conversely, suppose (5) holds. Then we may write (why?)

$$f(y) = \sup_{x \in \mathbb{R}^d} \{f(x) + \langle \nabla f(x), y - x\rangle\}$$

for any $y \in \mathbb{R}^d$. Since a pointwise supremum of an arbitrary collection of convex functions is convex, the function f must be convex. □

Examples

The following univariate functions are convex (check this!):

1. (Boltzmann-Shannon entropy)

$$f(x) = \begin{cases} x \log x & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ +\infty & \text{if } x < 0 \end{cases}$$

2. (Fermi-Dirac entropy)

$$f(x) = \begin{cases} x \log(x) + (1-x)\log(1-x) & \text{if } x \in (0,1) \\ 0 & \text{if } x \in \{0,1\} \\ +\infty & \text{otherwise} \end{cases}$$

3. (Hellinger)

$$f(x) = \begin{cases} -\sqrt{1-x^2} & \text{if } x \in [-1,1] \\ +\infty & \text{otherwise} \end{cases}$$

4. (Exponential) $f(x) = e^x$
5. (Log-exp) $f(x) = \log(1 + e^x)$

Polyak-Łojasiewicz inequality

Strongly convex functions satisfy the following useful property.

Lemma (PŁ-condition)

Any C^1-smooth and α-strongly convex function f satisfies

$$f(x) - f^* \leq \frac{1}{2\alpha}\|\nabla f(x)\|^2 \qquad \forall x.$$

Proof: Define the function

$$Q_x(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|^2.$$

Then we know

$$f(\bar{x}) \geq Q_x(\bar{x}) \geq \min_y Q_x(y) = f(x) - \frac{1}{2\alpha}\|\nabla f(x)\|^2.$$

Rearranging completes the proof. □

Gradient descent for smooth strongly convex functions

For any β-smooth and α-strongly convex function, the quotient

$$\boxed{\kappa = \frac{\beta}{\alpha}}$$

is called the condition number of f.

Theorem (Gradient descent under strong convexity)

Let f be an α-strongly convex and β-smooth function. Then the gradient descent iterates with $\eta = \frac{1}{\beta}$ satisfy

$$f(x_{t+1}) - f^* \leq \left(1 - \frac{1}{2\kappa}\right)(f(x_t) - f^*), \tag{6}$$

$$\|x_{t+1} - \bar{x}\|^2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)\|x_t - \bar{x}\|^2. \tag{7}$$

The linear rate is very sensitive to κ and in particular to small values of α.

Proof

The PL condition and the descent lemma yield

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2\beta}\|\nabla f(x_t)\|^2 \leq -\frac{1}{2\kappa}(f(x_t) - f^*).$$

Adding and subtracting f^* from both sides and rearranging gives (6).

Next, we prove (7). To this end, we successively compute

$$\begin{aligned}
\|x_{t+1} - \bar{x}\|^2 &= \|(x_t - \bar{x}) - \beta^{-1}\nabla f(x_t)\|^2 \\
&= \|x_t - \bar{x}\|^2 + \frac{2}{\beta}\langle \nabla f(x_t), \bar{x} - x_t\rangle + \frac{1}{\beta^2}\|\nabla f(x_t)\|^2 \\
&\leq \|x_t - \bar{x}\|^2 + \frac{2}{\beta}\left(f^* - f(x_t) - \frac{\alpha}{2}\|x_t - \bar{x}\|^2\right) + \frac{1}{\beta^2}\|\nabla f(x_t)\|^2 \\
&= \left(1 - \frac{\alpha}{\beta}\right)\|x_t - \bar{x}\|^2 + \frac{2}{\beta}\left(f^* - f(x_t) + \frac{1}{2\beta}\|\nabla f(x_t)\|^2\right). \qquad (8)
\end{aligned}$$

Namely, strong convexity and the descent lemma imply

$$f^* + \tfrac{\alpha}{2}\|x_{t+1} - \bar{x}\|^2 \leq f(x_{t+1}) \leq f(x_t) - \tfrac{1}{2\beta}\|\nabla f(x_t)\|^2,$$

and therefore

$$f^* - f(x_t) + \frac{1}{2\beta}\|\nabla f(x_t)\|^2 \leq -\frac{\alpha}{2}\|x_{t+1} - \bar{x}\|^2.$$

Combining this estimate with (8) and rearranging yields (7).

Sublinear rate for smooth and convex problems

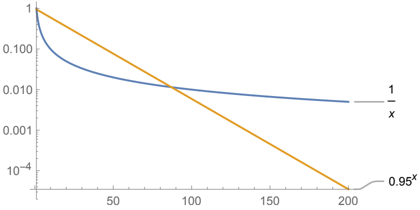
Theorem (Gradient descent under convexity)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and β-smooth function. Then the iterates generated by gradient descent with $\eta = \frac{1}{\beta}$ satisfy

$$f(x_t) - f^* \leq \frac{\beta \|x_0 - \bar{x}\|^2}{2t}.$$

Thus gradient descent satisfies the guarantee:

$$f(x_t) - f^* \leq \min\left\{\frac{1}{2t}, \left(1 - \frac{1}{2\kappa}\right)^t\right\} \cdot \beta \|x_0 - \bar{x}\|^2 \qquad \text{for all } t \geq 0.$$



Typically, the sublinear rate is observed in the early iterations of the algorithm, while the linear rate is observed towards the end (if at all).

Proof

Note that x_{t+1} is the minimizer of the β-strongly convex function

$$Q(y) = f(x_t) + \langle \nabla f(x_t), y - x_t \rangle + \tfrac{\beta}{2}\|y - x_t\|^2.$$

Therefore

$$\begin{aligned} f(x_{t+1}) &\le Q(x_{t+1}) \\ &\le Q(\bar{x}) - \frac{\beta}{2}\|x_{t+1} - \bar{x}\|^2 \\ &= f(x_t) + \langle \nabla f(x_t), \bar{x} - x_t \rangle + \tfrac{\beta}{2}\|x_t - \bar{x}\|^2 - \tfrac{\beta}{2}\|x_{t+1} - \bar{x}\|^2 \\ &\le f^* + \tfrac{\beta}{2}\left(\|x_t - \bar{x}\|^2 - \|x_{t+1} - \bar{x}\|^2\right). \end{aligned}$$

Subtracting f^* from both sides and summing, the terms on the right telescope:

$$\sum_{i=0}^{t-1}(f(x_{i+1}) - f^*) \le \tfrac{\beta}{2}\|x_0 - \bar{x}\|^2.$$

Since the values $\{f(x_i)\}_{i\ge 0}$ are nonincreasing, we deduce

$$f(x_t) - f^* \le \tfrac{1}{t}\sum_{i=0}^{t-1}(f(x_{i+1}) - f^*) \le \tfrac{\beta\|x_0 - \bar{x}\|^2}{2t},$$

as claimed. □

Accelerated gradient descent

Is there an algorithm that is guaranteed to succeed with fewer gradient evaluations? Yes!

Accelerated gradient method:

Initialization: $t = 0$ and $a_0 = a_{-1} = 1$, $x_{-1} = x_0$

For $t = 1, \ldots, T$ do

$$\left\{ \begin{aligned} u_t &= x_t + a_t(a_{t-1}^{-1} - 1)(x_t - x_{t-1}) \\ x_{t+1} &= u_t - \frac{1}{\beta}\nabla f(u_t) \\ a_{t+1} &= \frac{\sqrt{a_t^4 + 4a_t^2} - a_t^2}{2} \end{aligned} \right\}.$$

Theorem (Accelerated gradient method)

Let f be β-smooth and convex. Then the iterates generated by the accelerated gradient method satisfy

$$f(x_{t+1}) - f(x) \leq \frac{2\beta\|x_0 - x\|^2}{(t+2)^2} \qquad \forall x.$$

Proof

We will need the following basic lemma (check it!).

Lemma (Growth of a_t)

The following are true.

1. *The relation* $\frac{1-a_{t+1}}{a_{t+1}^2} = \frac{1}{a_t^2}$ *holds for all* $t \geq 0$.
2. *We have* $\sum_{i=0}^{t} \frac{1}{a_i} = \frac{1}{a_t^2}$ *and* $a_t \leq \frac{2}{t+2}$*, for each* $t \geq 0$.

Define $m_t(y) := f(u_t) + \langle \nabla f(u_t), y - u_t \rangle$ for each index t. Since x_{t+1} is the minimizer of the β-strongly convex function $m_t + \frac{\beta}{2}\|\cdot - u_t\|^2$, we estimate

$$
\begin{aligned}
f(x_{t+1}) &\leq m_t(x_{t+1}) + \frac{\beta}{2}\|x_{t+1} - u_t\|^2 \\
&\leq m_t(a_t x + (1-a_t)x_t) + \frac{\beta}{2}\|a_t x + (1-a_t)x_t - u_t\|^2 \\
&- \frac{\beta}{2}\|a_t x + (1-a_t)x_t - x_{t+1}\|^2 \\
&\leq a_t m_t(x) + (1-a_t)m_t(x_t) \\
&+ \frac{\beta a_t^2}{2}\left(\|x - [x_t - a_t^{-1}(x_t - u_t)]\|^2 - \|x - [x_t - a_t^{-1}(x_t - x_{t+1})]\|^2\right)
\end{aligned}
$$

Subtracting $f(x)$ from both sides and dividing by a_t^2 then yields

$$\begin{aligned}\frac{1}{a_t^2}(f(x_{t+1}) - \varphi(x)) \leq \frac{1-a_t}{a_t^2}(f(x_t) - f(x))& \\ + \frac{\beta}{2}\Big(\|x - [x_t - a_t^{-1}(x_t - u_t)]\|^2& \\ - \|x - [x_t - a_t^{-1}(x_t - x_{t+1})]\|^2\Big)&.\end{aligned} \tag{9}$$

The update rule for u_t makes the last two lines of (9) telescope. Indeed, define an auxiliary sequence $z_t = x_t - a_t^{-1}(x_t - u_t)$. Observe that z_{t+1} then satisfies

$$\begin{aligned}z_{t+1} = x_{t+1} - a_{t+1}^{-1}(x_{t+1} - u_{t+1}) &= x_{t+1} + (a_t^{-1} - 1)(x_{t+1} - x_t) \\ &= x_t - a_t^{-1}(x_t - x_{t+1}).\end{aligned}$$

Thus the inequality (9) becomes

$$\begin{aligned}\frac{1}{a_t^2}(f(x_{t+1}) - f(x)) + \frac{\beta}{2}\|x - z_{t+1}\|^2 &\leq \frac{1-a_t}{a_t^2}(f(x_t) - f(x)) + \frac{\beta}{2}\|x - z_t\|^2 \\ &= \frac{1}{a_{t-1}^2}(f(x_t) - f(x)) + \frac{\beta}{2}\|x - z_t\|^2,\end{aligned}$$

where the last equality uses the definition of a_t. Iterating the recurrence yields

$$\frac{1}{a_t^2}(f(x_{t+1}) - f(x)) \leq \frac{1-a_0}{a_0}(f(x_0) - f(x)) + \frac{\beta}{2}\|x - z_0\|^2.$$

thereby completing the proof. □

Summary

It is possible to modify the accelerated algorithm for β-smooth and α-strongly convex functions to have a rate of convergence

$$f(x_t) - f^\star \leq C(1 - \sqrt{\kappa})^t \|x_0 - \bar{x}\|^2.$$

for some numerical constant C. The following table summarizes our findings.

	Grad. Descent	Accelerated Grad. Descent
β-smooth and convex	$\frac{\beta\|x_0-\bar{x}\|^2}{\epsilon}$	$\sqrt{\frac{\beta\|x_0-\bar{x}\|^2}{\epsilon}}$
β-smooth and α-convex	$\frac{\beta}{\alpha}\log\left(\frac{f(x_0)-f^*}{\epsilon}\right)$	$\sqrt{\frac{\beta}{\alpha}}\log\left(\frac{\|x_0-\bar{x}\|^2}{\epsilon}\right)$

Table: Number of iterations t to reach $f(x_t) - f^* \leq \epsilon$

Subgradients

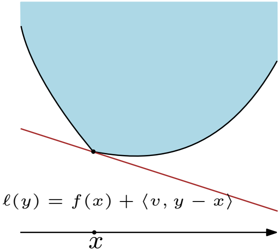
We will next look at algorithms for nonsmooth convex optimization.

Definition (Subdifferential)

Consider a convex function $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ and a point x, with $f(x)$ finite. Then a vector $v \in \mathbb{R}^d$ is called a subgradient of f at x if the inequality holds:

$$f(y) \geq f(x) + \langle v, y - x \rangle \qquad \forall y. \tag{10}$$

The set of all such vectors v is called the subdifferential *of f at x*, and is denoted by $\partial f(x)$. For points x at which $f(x)$ is infinite, we set $\partial f(x) = \emptyset$.



Calculus rules

Subdifferentials van be computed easily through the following calculus rule. For any convex function $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ define $\mathrm{dom} f = \{x : f(x) < \infty\}$.

Theorem (Calculus)

Let $f\colon \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ and $g\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be lower-semicontinuous convex functions, and let $A\colon \mathbb{R}^d \to \mathbb{R}^m$ be a linear map and $b \in \mathbb{R}^m$ a vector. Suppose the regularity condition

$$-b \in \mathrm{int}(\mathrm{dom}\ f) - A(\mathrm{int}(\mathrm{dom}\ g)).$$

Then the subdifferential of the function

$$h(x) = f(Ax - b) + g(x)$$

is given by

$$\partial h(x) = A^\top \partial f(Ax - b) + \partial g(x) \qquad \forall x.$$

The proof requires a bit of background and we will omit it.

Projections

It will be important for the problems we consider to work on constrained problems. We will incorporate a constraint set into algorithms through the nearest point projection. Along with any set $Q \subset \mathbb{R}^d$ define the distance

$$\mathrm{dist}_Q(y) := \inf_{x \in Q} \|x - y\|,$$

and the projection

$$\mathrm{proj}_Q(y) := \{x \in Q : \mathrm{dist}_Q(y) = \|x - y\|\}.$$

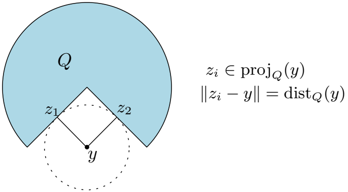


Figure: Nearest-point projection

Properties of projections

We will need the following basic theorem.

Theorem (Properties of the projection)

For any nonempty, closed, convex set $Q \subset \mathbb{R}^d$, the set $\operatorname{proj}_Q(y)$ is a singleton. Moreover, the closest point $z \in Q$ to y is characterized by the property:

$$\langle y - z, x - z \rangle \leq 0 \qquad \text{for all } x \in Q. \tag{11}$$

Consequently, the projection is 1-Lipschitz:

$$\|\operatorname{proj}_Q(y) - \operatorname{proj}_Q(x)\| \leq \|y - x\| \qquad \forall x, y.$$

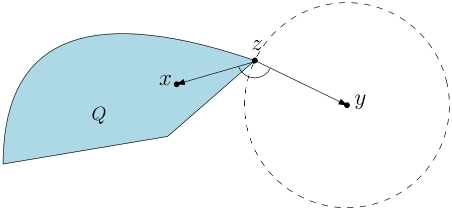


Figure: Nearest-point projection for convex sets

Proof

Fix a point $y \notin Q$. The claim that any point z satisfying (11) lies in $\mathrm{proj}_Q(y)$ is an easy exercise (verify it!). We therefore prove the converse. To this end, fix a point $z \in \mathrm{proj}_Q(y)$ and an arbitrary $x \in Q$. For each $t \in [0,1]$, define the point $x_t := z + t(x - z)$ and define the function $\varphi(t) := \frac{1}{2}\|y - x_t\|^2$. Convexity implies $x_t \in Q$ for all $t \in [0,1]$ and therefore

$$\varphi(t) \geq \tfrac{1}{2}\mathrm{dist}_Q^2(y) = \varphi(0).$$

Taking the derivative of φ, we therefore deduce

$$0 \leq \lim_{t \searrow 0} \frac{\varphi(t) - \varphi(0)}{t} = \varphi'(0) = -\langle y - z, x - z\rangle,$$

as claimed. Thus, a points z lies in $\mathrm{proj}_Q(y)$ if and only if (11) holds.

To see that $\mathrm{proj}_Q(y)$ is a singleton, consider any two points $z, z' \in \mathrm{proj}_Q(y)$. Then, the estimate (11) for z and z' (with $x = z'$ and $x = z$, respectively) becomes

$$\langle y - z, z' - z\rangle \leq 0 \qquad \text{and} \qquad \langle y - z', z - z'\rangle \leq 0.$$

Adding the two inequalities yields $0 \geq \langle z - z', z - z'\rangle = \|z - z'\|^2$, and therefore $z = z'$ as we had to show.

Proof (continued)

Now fix two point x and y and set $x^+ = \mathrm{proj}_Q(x)$ and $y^+ = \mathrm{proj}_Q(y)$.
Compute

$$\begin{aligned} \|x^+ - y^+\|^2 - \langle x^+ - y^+, x - y\rangle &= \langle x^+ - y^+, (x^+ - x) - (y^+ - y)\rangle \\ &= \underbrace{\langle y^+ - x^+, x - x^+\rangle}_{\le 0} + \underbrace{\langle x^+ - y^+, y - y^+\rangle}_{\le 0}. \end{aligned}$$

Rearranging and applying Cauchy–Schwarz inequality completes the proof. $\square$

Subgradient method

We now focus on the optimization problem

$$\min_{x \in Q} \; f(x), \tag{12}$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$ is a convex function that is L-Lipschitz continuous on a neighborhood of a closed convex set $Q \subset \mathbb{R}^d$. We let $\bar{x}$ denote the minimizer of the problem and set $f^* = f(\bar{x})$.

The projected subgradient method proceeds according to the rule:

For $t = 1, \ldots, T$ do

$$\left\{ \begin{matrix} \text{Choose } v_t \in \partial f(x_t) \\ \text{Set } x_{t+1} = \mathrm{proj}_Q(x_t - \eta_t v_t) \end{matrix} \right\}.$$

where $\eta_t > 0$ are to be chosen.

Subgradient method under convexity

Theorem (Subgradient method under convexity)

Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a convex function that is L-Lipschitz continuous on a neighborhood of a closed convex set $Q \subset \mathbb{R}^d$. Then the iterates satisfy

$$f\left(\frac{1}{\sum_{i=0}^t \eta_i}\sum_{i=0}^{t}\eta_i x_i\right) - f^* \leq \frac{\|x_0 - \bar{x}\|^2 + L^2\sum_{i=0}^t \eta_i^2}{2\sum_{i=0}^t \eta_i}. \qquad (13)$$

In particular, when using the constant parameter $\eta_t = \frac{R}{L\sqrt{T+1}}$ for a fixed $R \geq \|x_0 - \bar{x}\|$, the efficiency estimate becomes

$$f\left(\frac{1}{T+1}\sum_{t=0}^{T} x_t\right) - f^* \leq \frac{RL}{\sqrt{T+1}}. \qquad (14)$$

Proof

We successively compute

$$\begin{aligned} \|x_{t+1} - \bar{x}\|^2 &= \|\mathrm{proj}_Q(x_t - \eta_t v_t) - \bar{x}\|^2 \\ &= \|\mathrm{proj}_Q(x_t - \eta_t v_t) - \mathrm{proj}_Q(\bar{x})\|^2 \\ &\leq \|(x_t - \bar{x}) - \eta_t v_t\|^2 && (15) \\ &= \|x_t - \bar{x}\|^2 - 2\eta_t \langle v_t, x_t - \bar{x}\rangle + \eta_t^2 \|v_t\|^2, && (16) \\ &\leq \|x_t - \bar{x}\|^2 - 2\eta_t (f(x_t) - f^*) + \eta_t^2 L^2, && (17) \end{aligned}$$

where (23) uses that proj_Q is 1-Lipschitz continuous and (17) uses convexity and Lipschitz continuity of f. Iterating the recursion yields

$$\|x_{T+1} - \bar{x}\|^2 \leq \|x_0 - x^*\|^2 - 2\sum_{t=0}^{T} \eta_t (f(x_t) - f^*) + L^2 \sum_{t=0}^{T} \eta_t^2.$$

Lower-bounding the left side by zero and rearranging, we conclude

$$\sum_{t=0}^{T} \eta_t (f(x_t) - f^*) \leq \frac{\|x_0 - \bar{x}\|^2 + L^2 \sum_{t=0}^{T} \eta_t^2}{2}. \qquad (18)$$

Proof continued

Finally using convexity, observe

$$f\left(\frac{1}{\sum_{t=0}^{T}\eta_t}\sum_{t=0}^{T}\eta_t x_t\right) - f^* \leq \frac{\sum_{t=0}^{T}\eta_t(f(x_t)-f^*)}{\sum_{i=0}^{t}\eta_t}.$$

Combining this estimate with (25) completes the proof of (13). Setting $\eta_t = \eta$ for all $t = 0, \ldots, T-1$ in (13) yields the guarantee

$$f\left(\frac{1}{T+1}\sum_{t=0}^{T}x_t\right) - f^* \leq \frac{\|x_0 - x^*\|^2}{2(T+1)\eta} + \frac{L^2\eta}{2}.$$

Optimizing the right side of (13) in η yields the choice $\eta = \frac{R}{L\sqrt{T+1}}$ and the guarantee (14). □

Subgradient method under strong convexity

A faster convergence rate is possible under strong convexity.

Theorem (Subgradient method under strong convexity)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be an α-strongly convex function that is L-Lipschitz continuous on a neighborhood of a closed convex set $Q \subset \mathbb{R}^d$. Then the iterates with $\eta_t = \frac{2}{\alpha(t+1)}$ satisfy

$$f\left(\frac{2}{t(t+1)}\sum_{i=1}^{t} i x_i\right) - f^* \leq \frac{2L^2}{\alpha(t+1)}.$$

Proof

From (16) and Lipschitz continuity and strong convexity of f, we compute

$$\begin{aligned}\|x_{t+1}-\bar{x}\|^2 &\leq \|x_t-\bar{x}\|^2 + 2\eta_t\langle v_t, \bar{x}-x_t\rangle + \eta_t^2\|v_t\|^2 \\ &\leq \|x_t-\bar{x}\|^2 + 2\eta_t\Big(f^* - f(x_t) - \tfrac{\alpha}{2}\|x^*-x_t\|^2\Big) + \eta_t^2 L^2.\end{aligned}$$

Rearranging and diving through by $2\eta_t$ yields the expression

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

Plugging in $\eta_t := \frac{2}{\alpha(t+1)}$ and multiplying through by t, we obtain

$$t\Big(f(x_t)-f(\bar{x})\Big) \leq \frac{\alpha t(t-1)}{4}\|x_t-x^*\|^2 - \frac{\alpha t(t+1)}{4}\|x_{t+1}-\bar{x}\|^2 + \frac{t}{\alpha(t+1)}L^2.$$

Summing for $i = 1\ldots,t$, the first two terms on the right telescope, yielding

$$\sum_{i=1}^{t} i\Big(f(x_i)-f(\bar{x})\Big) \leq \sum_{i=1}^{t} \tfrac{i}{\alpha(i+1)}L^2 \leq \frac{tL^2}{\alpha}.$$

Dividing through by $\sum_{i=1}^t i = \frac{t(t+1)}{2}$ and using convexity of f we conclude

$$f\left(\tfrac{2}{t(t+1)}\sum_{i=1}^{t} i x_i\right) - f^* \leq \left(\frac{1}{\sum_{i=1}^t i}\right)\cdot\sum_{i=1}^{t} i\Big(f(x_i)-f(\bar{x})\Big) \leq \frac{2L^2}{\alpha(t+1)},$$

as claimed. □

Lower bounds for convex optimization

Summary of what we have so far:

	convex, β-smooth	α-strongly convex, β-smooth
Gradient descent	$\frac{\beta\|x_0 - x^*\|^2}{\epsilon}$	$\kappa \cdot \log(\frac{f(x_0) - f^*}{\epsilon})$
Accel. grad. descent	$\sqrt{\frac{\beta\|x_0 - x^*\|^2}{\epsilon}}$	$\sqrt{\kappa} \cdot \log(\frac{f(x_0) - f^*}{\epsilon})$

Table: Number of gradient evaluations to find x satisfying $f(x) - f^* \leq \epsilon$

	convex, L-Lipschitz	α-strongly convex, L-Lipschitz
Subgrad. method	$\frac{L^2 R^2}{\epsilon^2}$	$\frac{L^2}{\alpha\epsilon}$

Table: Number of subgradient evaluations to find x satisfying $f(x) - f^* \leq \epsilon$, where an upper bound $R \geq \|x_0 - x^*\|$ is assumed to be known.

We will next see that the accelerated gradient method is minimax optimal for smooth minimization and the subgrdient method is minimax optimal for nonsmooth optimization. We omit all proofs since they are quite tedious.

Lower bounds for convex optimization

We will focus on the problem $\min_{x\in\mathbb{R}^d} f(x)$. The algorithms we consider access information about f by querying a "first-order oracle", which on input $x \in \mathbb{R}^d$ returns some subgradient $v \in \partial f(x)$. We will prove lower-complexity bounds for a large class of algorithms, summarized in the following definition.

Definition (Linearly-expanding first-order method)

An algorithm is called a linearly-expanding first-order method if it generates an iterate sequence $\{x_k\}$ satisfying

$$x_t \in x_0 + \text{span}\{v_0, \ldots, v_{t-1}\} \qquad \text{for } t \geq 1,$$

where $v_i \in \partial f(x_i)$ is generated by a first-order oracle of f with input x_i.

The lower-bounds that appear next hold for a wider class of algorithms, but the statements become more cumbersome.

Lower bounds for convex optimization

Theorem (Lower-complexity bound for smooth convex optimization)

Fix a dimension $d \in \mathbb{N}$, a counter $1 \leq t \leq (n-1)/2$, and a constant $\beta > 0$. Then there exists a convex β-smooth function $f\colon \mathbb{R}^d \to \mathbb{R}$ so that the iterates generated by any linearly-expanding first-order method started at x_0 satisfy

$$f(x_t) - \min f \geq \frac{3\beta \|x_0 - \bar{x}\|^2}{32(t+1)^2}, \tag{19}$$

where x^* is any minimizer of f.

An entirely analogous statement holds for α-strongly convex and β-smooth functions with the lower-complexity bound becoming

$$f(x_t) - f^* \geq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2t} \|x_0 - \bar{x}\|^2. \tag{20}$$

Lower bounds for convex optimization

Theorem (Lower-complexity bound for nonsmooth convex optimization)

Fix a dimension $d \in \mathbb{N}$, an iteration counter $t \leq d$, and a real $L > 0$. Then there exists a convex function $f : \mathbb{R}^d \to \mathbb{R}$ that is L-Lipschitz continuous on a ball $B_R(0)$, for some $R > 0$, and such that any linear expanding first-order method initialized at the origin satisfies

$$\min_{k=1,\ldots,t-1} f(x_k) - \min_{x \in B_R(0)} f(x) \geq \frac{RL}{2(1+\sqrt{t})}.$$

An entirely analogous statement holds for α-strongly convex and L-Lipschitz functions on $B_R(0)$ with the lower-complexity bound becoming

$$\min_{k=1,\ldots,t-1} f(x_k) - \min_{x \in B_R(0)} f(x) \geq \frac{L^2}{8\alpha t}$$

Conclusion: There is a huge gap between efficiency of algorithms for smooth optimization and nonsmooth optimization: $O(\frac{1}{\sqrt{\epsilon}})$ vs $O(\frac{1}{\epsilon^2})$. We will later see nonsmooth problems that are highly structured and algorithms that use this structure have rates that are close to that for smooth optimization.

Stochastic gradient for least squares

Problem:

$$\min_x \; f(x) = \frac{1}{2} \mathop{\mathbb{E}}_{(a,b)\sim\mathcal{P}} (a^\top x - b)^2,$$

where $b = \langle a, \bar{x} \rangle + \epsilon$ for some fixed $\bar{x} \in \mathbb{R}^d$ and random noise ϵ_i.

Stochastic gradient method (Online Least Squares):

$$\begin{Bmatrix} \text{Draw } (a_t, b_t) \sim \mathcal{P} \\ \text{Set } x_{t+1} = x_t - \eta_t (a_t^\top x_t - b_t) a_t \end{Bmatrix}.$$

Throughout $\mathbb{E}_t = \mathbb{E}[\cdot \mid x_t]$ will denote the conditional expectation.

Theorem (One step improvement)

Define the covariance matrix $\Sigma := \mathbb{E} a a^\top$ *and suppose:*

$$\mathbb{E}[\epsilon \mid a] = 0, \qquad \mathbb{E}[\epsilon \mid a] \leq \sigma^2, \qquad \alpha I \preceq \Sigma, \qquad \mathbb{E}[a a^\top \|a\|^2] \preceq R^2 \Sigma.$$

Then it holds:

$$\mathbb{E}_t \|x_{t+1} - \bar{x}\|^2 \leq (1 - \alpha \eta_t (2 - \eta_t R^2)) \|x_t - \bar{x}\|^2 + \eta_t^2 \sigma^2 \mathrm{tr}(\Sigma).$$

Proof

We compute

$$\begin{aligned}\|x_{t+1}-\bar{x}\|^2 &= \|(x_t-\bar{x})-\eta_t(a_t^\top x_t - b_t)a_t\|^2 \\ &= \|x_t-\bar{x}\| - 2\eta_t \underbrace{\langle (a_t^\top x_t - b_t)a_t, x_t-\bar{x}\rangle}_{P_1} + \eta_t^2 \underbrace{\|(a_t^\top x_t - b_t)a_t\|^2}_{P_2}.\end{aligned}$$

Taking the conditional expectation yields

$$\mathbb{E}[P_1 \mid a_t, x_t] = \langle (a_t^\top x_t - a_t^\top \bar{x})a_t, x_t - \bar{x}\rangle = (a_t^\top (x_t-\bar{x}))^2$$

and

$$\begin{aligned}\mathbb{E}[P_2 \mid a_t, x_t] &= \mathbb{E}[(a_t^\top x_t - b_t)^2 \mid a_t, x_t]\cdot \|a_t\|^2 \\ &= (a_t^\top (x_t-\bar{x}))^2 \cdot \|a_t\|^2 + \mathbb{E}[\epsilon^2 \mid a_t]\cdot \|a_t\|^2.\end{aligned}$$

Taking expectation now with respect to a_t, get

$$\mathbb{E}[P_1 \mid x_t] = \|x_t-\bar{x}\|_\Sigma^2, \qquad \mathbb{E}[P_2 \mid x_t] \le R^2\|x_t-\bar{x}\|_\Sigma^2 + \sigma^2 \mathrm{tr}(\Sigma)$$

Thus we conclude

$$\mathbb{E}_t\|x_{t+1}-\bar{x}\|^2 \le (1-\alpha\eta_t(2-\eta_t R^2))\|x_t-\bar{x}\|^2 + \eta_t^2\sigma^2\mathrm{tr}(\Sigma),$$

as claimed. □

Stochastic gradient for least squares

After unrolling the recursion, one possible choice of η_t is on the order of $1/t$. The resulting convergence rate becomes the following.

Theorem (Convergence rate)

Set $\eta_t = \frac{2}{\alpha t + 2R^2}$. *Then the iterates* x_t *satisfy*

$$\mathbb{E}\|x_t - \bar{x}\|^2 \leq \frac{\max\{\alpha^2(1+\frac{2R^2}{\alpha})\|x_1 - \bar{x}\|^2, 4\sigma^2 \mathrm{tr}(\Sigma)\}}{\alpha^2(t+\frac{2R^2}{\alpha})}$$

Thus, the rate is roughly

$$\mathbb{E}\|x_t - \bar{x}\|^2 = O\left(\frac{\sigma^2 \mathrm{tr}(\Sigma)}{\alpha^2 t}\right).$$

This rate is suboptimal in a number of ways. Looking at the Le Cam's asymptotic lower bound, we would expect a rate on the order of $O\left(\frac{\sigma^2 \mathrm{tr}(\Sigma^{-1})}{t}\right)$. Similarly, we expect the function gap to be on the order of $\mathbb{E}\|x_t - \bar{x}\|_\Sigma = O(\frac{\sigma^2 d}{n})$. It turns out these estimates are not achieved by x_t but are achieved by the average iterate $\hat{x}_t = \frac{1}{t}\sum_{i=1}^t x_i$. We will not prove this fact, but will see the important role of averaging more generally.

Proof

Taking expectation with respect to $a_1, \ldots, x_t$ and using the tower-rule we get

$$\begin{aligned}\mathbb{E}\|x_{t+1} - \bar{x}\|^2 &\leq (1 - \alpha\eta_t(2 - \eta_t R^2))\mathbb{E}\|x_t - \bar{x}\|^2 + \eta_t^2\sigma^2\mathrm{tr}(\Sigma) \\ &\leq \left(1 - \frac{2}{t + 2R^2/\alpha}\right)\mathbb{E}\|x_t - \bar{x}\|^2 + \frac{4\sigma^2\mathrm{tr}(\Sigma)/\alpha^2}{(t + 2R^2/\alpha)^2}\end{aligned}$$

We can now use the following elementary lemma on convergence of sequences, which can be quickly proved by induction (do it!).

Lemma: Consider a sequence $D_t > 0$ and constants $t_0 \geq 0$, $a > 0$ satisfying

$$D_{t+1} \leq (1 - \frac{2}{t + t_0})D_t + \frac{a}{(t + t_0)^2}.$$

Then the estimate $D_t \leq \frac{\max\{(1+t_0)D_1, a\}}{t+t_0}$ holds for all t.

Setting $t_0 = \frac{2R^2}{\alpha}$ and $a = 4\sigma^2\mathrm{tr}(\Sigma)/\alpha^2$ completes the proof. □

Stochastic gradient method for convex problems

Problem:

$$\min_{x \in Q} \; f(x)$$

where Q is closed and convex and f is convex and L-Lipschitz on a neighborhood of Q.

Stochastic gradient oracle: Suppose that there exists a probability space $(\mathcal{Z}, \mathcal{F}, \mathcal{P})$ and a measurable map $G\colon \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}^d$ satisfying

$$\mathbb{E}_z[G(x,z)] \in \partial f(x) \qquad \text{and} \qquad \mathbb{E}_z\|G(x,z)\|^2 \leq L \qquad \forall x \in Q.$$

Main example is $G(x,z) = \nabla \ell(x,z)$ or $G(x,(z_1,\ldots,z_k)) = \frac{1}{k}\sum_{i=1}^k \nabla \ell(x, z_i)$.

Remark: Many variants of stochastic gradient oracles are possible.

Projected stochastic gradient method:

$$\begin{Bmatrix} \text{Draw } z_k \sim \mathcal{P} \\ \text{Set } x_{t+1} = \mathrm{proj}_Q(x_t - \eta_t G(x_t, z_t)) \end{Bmatrix}.$$

Stochastic gradient method for convex problems

Theorem (Stochastic subgradient method under convexity)

Suppose that f is convex and L-Lipschitz on a neighborhood of a closed convex set Q. Then the iterates x_t satisfy

$$\mathbb{E}f\left(\frac{1}{\sum_{i=0}^{t}\eta_i}\sum_{i=0}^{t}\eta_i x_i\right) - f^* \leq \frac{\|x_0-\bar{x}\|^2 + L^2\sum_{i=0}^{t}\eta_i^2}{2\sum_{i=0}^{t}\eta_i}. \tag{21}$$

In particular, when using the constant parameter $\eta_t = \frac{R}{L\sqrt{T+1}}$ for a fixed $R \geq \|x_0 - \bar{x}\|$, the efficiency estimate becomes

$$\mathbb{E}f\left(\frac{1}{T+1}\sum_{t=0}^{T} x_t\right) - f^* \leq \frac{RL}{\sqrt{T+1}}. \tag{22}$$

Proof

Set $v_t := G(x_t, z_t)$. We successively compute

$$\begin{aligned}
\|x_{t+1} - \bar{x}\|^2 &= \|\mathrm{proj}_Q(x_t - \eta_t v_t) - \bar{x}\|^2 \\
&= \|\mathrm{proj}_Q(x_t - \eta_t v_t) - \mathrm{proj}_Q(\bar{x})\|^2 \\
&\leq \|(x_t - \bar{x}) - \eta_t v_t\|^2 \qquad (23) \\
&= \|x_t - \bar{x}\|^2 - 2\eta_t \langle v_t, x_t - \bar{x}\rangle + \eta_t^2 \|v_t\|^2,
\end{aligned}$$

where (23) uses that proj_Q is 1-Lipschitz continuous. Taking conditional expectation $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid x_t]$, we compute

$$\begin{aligned}
\mathbb{E}_t\|x_{t+1} - \bar{x}\|^2 &= \|x_t - \bar{x}\|^2 - 2\eta_t \langle \mathbb{E}_z G(x_t, z), x_t - \bar{x}\rangle + \eta_t^2 \mathbb{E}_z \|G(x_t, z)\|^2 \\
&\leq \|x_t - \bar{x}\|^2 - 2\eta_t (f(x_t) - f^*) + \eta_t^2 L^2, \qquad (24)
\end{aligned}$$

where (24) uses convexity of f and the definition of the stochastic subgradient oracle. Taking now expectation of both sides with respect to x_t and using the tower rule we deduce

$$\mathbb{E}\|x_{t+1} - \bar{x}\|^2 \leq \mathbb{E}\|x_t - \bar{x}\|^2 - 2\eta_t \mathbb{E}(f(x_t) - f^*) + \eta_t^2 L^2.$$

Proof continued

Iterating the recursion yields

$$\mathbb{E}\|x_{T+1} - \bar{x}\|^2 \leq \|x_0 - x^*\|^2 - 2\sum_{t=0}^{T} \eta_t \mathbb{E}(f(x_t) - f^*) + L^2 \sum_{t=0}^{T} \eta_t^2.$$

Lower-bounding the left side by zero and rearranging, we conclude

$$\sum_{t=0}^{T} \eta_t \mathbb{E}(f(x_t) - f^*) \leq \frac{\|x_0 - \bar{x}\|^2 + L^2 \sum_{t=0}^{T} \eta_t^2}{2}. \tag{25}$$

Finally using convexity, observe

$$\mathbb{E}f\left(\frac{1}{\sum_{t=0}^{T} \eta_t} \sum_{t=0}^{T} \eta_t x_t\right) - f^* \leq \frac{\sum_{t=0}^{T} \eta_t (\mathbb{E}f(x_t) - f^*)}{\sum_{i=0}^{t} \eta_t}.$$

Combining this estimate with (25) completes the proof of (21). Setting $\eta_t = \eta$ for all $t = 0, \ldots, T-1$ in (21) yields the guarantee

$$\mathbb{E}f\left(\frac{1}{T+1} \sum_{t=0}^{T} x_t\right) - f^* \leq \frac{\|x_0 - x^*\|^2}{2(T+1)\eta} + \frac{L^2 \eta}{2}.$$

Optimizing the right side in η yields $\eta = \frac{R}{L\sqrt{T+1}}$ and the guarantee (22). $\square$

Stochastic subgradient method under strong convexity

A faster convergence rate is possible under strong convexity.

Theorem (Stochastic subgradient method under strong convexity)

Suppose that f is α-convex and L-Lipschitz on a neighborhood of a closed convex set Q. Then the iterates x_t with $\eta_t = \frac{2}{\alpha(t+1)}$ satisfy

$$\mathbb{E} f\left(\frac{2}{t(t+1)} \sum_{i=1}^{t} i x_i\right) - f^* \leq \frac{2L^2}{\alpha(t+1)}.$$

Proof

The same argument as leading to (24), but now using strong convexity, gives

$$\mathbb{E}_t\|x_{t+1}-\bar{x}\|^2 \le \|x_t-\bar{x}\|^2 + 2\eta_t\Big(f^* - f(x_t) - \tfrac{\alpha}{2}\|x^*-x_t\|^2\Big) + \eta_t^2 L^2.$$

Rearranging, taking expectation in x_t, and using the tower rule yields

$$\mathbb{E}f(x_t) - f^* \le \left(\frac{1-\alpha\eta_t}{2\eta_t}\right)\mathbb{E}\|x_t-\bar{x}\|_2^2 - \frac{1}{2\eta_t}\mathbb{E}\|x_{t+1}-\bar{x}\|_2^2 + \frac{\eta_t}{2}L^2.$$

Plugging in $\eta_t := \frac{2}{\alpha(t+1)}$ and multiplying through by t, we obtain

$$t\Big(\mathbb{E}f(x_t) - f(\bar{x})\Big) \le \frac{\alpha t(t-1)}{4}\mathbb{E}\|x_t - x^*\|^2 - \frac{\alpha t(t+1)}{4}\mathbb{E}\|x_{t+1}-\bar{x}\|^2 + \frac{t}{\alpha(t+1)}L^2.$$

Summing for $i = 1\dots,t$, the first two terms on the right telescope, yielding

$$\sum_{i=1}^t i\Big(\mathbb{E}f(x_i) - f(\bar{x})\Big) \le \sum_{i=1}^t \tfrac{i}{\alpha(i+1)}L^2 \le \frac{tL^2}{\alpha}.$$

Dividing through by $\sum_{i=1}^t i = \frac{t(t+1)}{2}$ and using convexity of f we conclude

$$\mathbb{E}f\left(\tfrac{2}{t(t+1)}\sum_{i=1}^t i x_i\right) - f^* \le \left(\frac{1}{\sum_{i=1}^t i}\right)\cdot\sum_{i=1}^t i\Big(\mathbb{E}f(x_i) - f(\bar{x})\Big) \le \frac{2L^2}{\alpha(t+1)},$$

as claimed. □

Polyak-Juditsky averaging

As can be seen from the previous theorem, averaging gradients is important. In fact, the following theorem (stated informally) shows that averaging leads to an asymptotically optimal algorithm for stochastic optimization. We will omit the proof since it is quite technical.

Theorem (Polyak-Juditsky '92 (informal))

Consider minimizing $f(x) = \mathbb{E}_{z\sim\mathcal{P}}\ell(x,z)$ over $\mathbb{R}^d$ and let $\bar{x}$ be a minimizer of f satisfying $\nabla^2 f(\bar{x}) \succ 0$. Let x_t be the iterates generated by the stochastic gradient method with $\eta_t = \eta_0 t^{-\gamma}$ for some $\gamma \in (0.5, 1)$. Then under mild moment assumptions, the iterates x_t converge to $\bar{x}$ almost surely and the average iterate $\hat{x}_t = \frac{1}{t}\sum_{i=1}^t x_i$ satisfies

$$\sqrt{t}(\hat{x}_t - \bar{x}) \xrightarrow{d} \mathsf{N}\Big(0, \nabla^2 f(\bar{x})^{-1} \cdot \mathrm{Cov}(\nabla f(\bar{x}, z)) \cdot \nabla^2 f(\bar{x})^{-1}\Big).$$

Conclusion:

- Asymptotics of $\hat{x}_t$ match those of the sample average approximation.
- The average iterate $\hat{x}_t$ converges at a $t^{-1/2}$ rate regardless of choice of γ.

Stochastic Variance Reduced Gradient

Recall that **empirical risk minimization** is a problem of the form:

$$\min_{x \in \mathbb{R}^d} \; f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x).$$

The (sug)gradient algorithms we have considered used a single gradient evaluation $\nabla f(x)$ in each iteration. Evaluating $\nabla f(x)$ in principle requires evaluating n individual gradients $\nabla f_i(x)$, which is very expensive when n is large. Let us therefore instead think of evaluating $\nabla f_i(x)$ as a single unit of cost. Then the complexity of gradient descent becomes $O(n\kappa \log(\frac{f(x_0)-f^*}{\epsilon}))$.

We will now show that there exists an algorithm with the much better complexity $O\left((n+\kappa)\log(\frac{f(x_0)-f^*}{\epsilon})\right)$.

Remark: There are a few algorithms the achieve this improved rate (each having some advantages). We will focus on just one of them called Stochastic Variance Reduced Gradient (SVRG).

Variance reduction

Assumption: Each f_i is β-smooth and convex, and f is α-strongly convex.

Let us look at an algorithm with an update of the form

$$\boxed{x_{t+1} = x_t - \eta v_t},$$

where v_t is a random vector to be specified. As usual, we may write

$$\|x_{t+1} - \bar{x}\|^2 = \|x_t - \bar{x}\|^2 - 2\eta\langle v_t, x_t - \bar{x}\rangle + \eta^2\|v_t\|^2.$$

As long as $\mathbb{E}_t[v_t] = \nabla f(x_t)$, we may take expectations and obtain

$$\mathbb{E}\|x_{t+1} - \bar{x}\|^2 \leq \mathbb{E}\|x_t - \bar{x}\|^2 - 2\eta(f(x_t) - f^*) + \eta^2\mathbb{E}\|v_t\|^2. \tag{26}$$

In order to reach ϵ-accuracy, we must shrink η inversely to $\mathbb{E}\|v_t\|^2$. In order to allow larger stepsizes, we can aim to design a random unbiased stochastic gradient estimator with small variance.

Variance reduction

Here is one conceptually simple choice:

$$v_t = \nabla f_{i_t}(x_t) - \nabla f_{i_t}(\bar{x}),$$

where i_t is drawn uniformly at random from $\{1, \ldots, n\}$. Since we do not know $\bar{x}$, this vector is not computable directly but it does have a small variance. To see this, we can use the following lemma.

Lemma

Any β-smooth function $g\colon \mathbb{R}^d \to \mathbb{R}$ satisfies

$$\tfrac{1}{2\beta}\|\nabla g(x) - \nabla g(y)\|^2 \leq g(y) - g(x) - \langle \nabla g(x), y - x\rangle \qquad \forall x, y$$

Proof: Invoke descent $0 \leq Q(y - \beta^{-1}\nabla Q(y)) \leq Q(y) - \frac{1}{2\beta}\|\nabla Q(y)\|^2$ for $Q(y) := g(y) - g(x) - \langle \nabla g(x), y - x\rangle$. $\square$

Applying the lemma to each f_{i_t} yields

$$\boxed{\mathbb{E}_t v_t = \nabla f(x_t) \qquad \text{and} \qquad \mathbb{E}_t\|v_t\|^2 \leq 2\beta(f(x_t) - f^*)}.$$

The second moment tends to zero along the iterates!

Variance reduction

Since we do not know $\bar{x}$, suppose instead that we have an approximate minimizer y and form the SVRG estimator

$$\boxed{v_t = \nabla f_{i_t}(x_t) - \nabla f_{i_t}(y) + \nabla f(y)}.$$

Then clearly $\mathbb{E}_t v_t = \nabla f(x_t)$ and we may estimate the variance

$$\begin{aligned}
\mathbb{E}_t\|v_t\|^2 &\le 2\mathbb{E}_t\|\nabla f_{i_t}(x_t) - \nabla f_{i_t}(\bar{x})\|^2 + 2\mathbb{E}_t\|\nabla f_{i_t}(\bar{x}) - \nabla f_{i_t}(y) + \nabla f(y)\|^2 \\
&\le 4\beta(f(x_t) - f^*) + 2\mathbb{E}_t\|\nabla f_{i_t}(\bar{x}) - \nabla f_{i_t}(y)\|^2 \\
&\le 4\beta(f(x_t) - f^*) + 4\beta(f(y) - f^*),
\end{aligned}$$

where the second and third inequalities follow from the lemma.

Let us now initialize $x_1 = y$ and see how many iterations are required to drive the gap $f(x_t) - f^*$ below a fraction of $f(x_1) - f^*$.

Variance reduction

Observe (26) becomes

$$\mathbb{E}\|x_{t+1} - \bar{x}\|^2 \leq \mathbb{E}\|x_1 - \bar{x}\|^2 - 2\eta(1 - 2\beta\eta)(f(x_t) - f^*) + 4\beta\eta^2(f(y) - f^*).$$

Iterating gives

$$\mathbb{E}\|x_{t+1} - \bar{x}\|^2 \leq \mathbb{E}\|y - \bar{x}\|^2 - 2\eta(1 - 2\beta\eta)\sum_{i=1}^{t}\mathbb{E}(f(x_i) - f^*) + 4\beta\eta^2 t(f(y) - f^*).$$

Lower bounding the left side by zero and noting $\frac{\alpha}{2}\|y - \bar{x}\|^2 \leq f(y) - f^*$ gives

$$\mathbb{E}f\left(\frac{1}{t}\sum_{i=1}^{t} x_i\right) - f^* \leq \left(\frac{1}{\alpha\eta(1 - 2\beta\eta)t} + \frac{2\beta\eta}{1 - 2\beta\eta}\right)(f(y) - f^*).$$

Setting $y^+ := \frac{1}{t}\sum_{i=1}^{t} x_i$, $\eta = \frac{1}{10\beta}$, and $t = 20\beta/\alpha$ we deduce

$$\boxed{\mathbb{E}f(y^+) - f^* \leq 0.9(f(y) - f^*)}.$$

Variance reduction

Thus in $t = \frac{\beta}{\alpha}$ iterations, the method shrinks the suboptimality gap by a constant fraction. The SVRG algorithm simply repeats this process in epochs. The cost of each epoch is one computation of the full gradient $\nabla f(y)$ and t computations of the individual gradients $\nabla f_i(x)$. Thus the method will find a point y satisfying $\mathbb{E}f(y) - f^* \leq \epsilon$ after having computed at most

$$O\left(\left(n + \frac{\beta}{\alpha}\right)\log\left(\frac{f(x_1) - f^*}{\epsilon}\right)\right),$$

individual gradients $\nabla f_i(x)$.