Chapter 4
Optimization

Ernest K. Ryu
Seoul National University

Mathematical Machine Learning Theory
Spring 2024

"...in fact, the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity."

— R. Tyrrell Rockafellar, *SIAM Review*, 1993 —

Risk decomposition

Let

$$\theta^\star \in \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \mathcal{R}[f_\theta], \qquad \hat{\theta}^\star \in \operatorname*{argmin}_{\hat{\theta} \in \mathbb{R}^p} \hat{\mathcal{R}}[f_{\hat{\theta}}].$$

Then,

$$\begin{aligned}
\mathcal{R}[f_{\hat{\theta}}] - \mathcal{R}^\star &= \underbrace{(\mathcal{R}[f_{\hat{\theta}}] - \hat{\mathcal{R}}[f_{\hat{\theta}}])}_{=\text{Estimation error}} + \underbrace{(\hat{\mathcal{R}}[f_{\theta^\star}] - \mathcal{R}[f_{\theta^\star}])}_{=\text{Estimation error}} \\
&\quad \underbrace{(\mathcal{R}[f_{\theta^\star}] - \mathcal{R}^\star)}_{=\text{Approximation error}} + \underbrace{(\hat{\mathcal{R}}[f_{\hat{\theta}}] - \hat{\mathcal{R}}[f_{\theta^\star}])}_{=\text{Optimization error}} \\
&\leq (\mathcal{R}[f_{\hat{\theta}}] - \hat{\mathcal{R}}[f_{\hat{\theta}}]) + (\hat{\mathcal{R}}[f_{\theta^\star}] - \mathcal{R}[f_{\theta^\star}]) \\
&\quad (\mathcal{R}[f_{\theta^\star}] - \mathcal{R}^\star) + \underbrace{(\hat{\mathcal{R}}[f_{\hat{\theta}}] - \hat{\mathcal{R}}[f_{\hat{\theta}^\star}])}_{=\text{Optimization error} \geq 0}.
\end{aligned}$$

We now discuss algorithms for solving

$$\operatorname*{minimize}_{\hat{\theta} \in \mathbb{R}^d} \quad \hat{\mathcal{R}}[f_{\hat{\theta}}].$$

Gradient descent

Consider the optimization problem

$$\underset{\theta \in \mathbb{R}^d}{\text{minimize}} \quad F(\theta)$$

where $F \colon \mathbb{R}^d \to \mathbb{R}$ is differentiable.

We consider *gradient descent*

$$\theta^{k+1} = \theta^k - \alpha_k \nabla F(\theta^k)$$

where $\theta^0 \in \mathbb{R}^d$ is a starting point and $\alpha_0, \alpha_1, \ldots \in \mathbb{R}$ is a positive sequence of stepsizes.

Outline

Least-squares to standard quadratic form

Consider the least-squares problem

$$\underset{\theta \in \mathbb{R}^d}{\text{minimize}} \quad \tfrac{1}{2}\|\Phi\theta - Y\|^2,$$

where $\Phi \in \mathbb{R}^{N \times d}$ and $Y \in \mathbb{R}^d$. Let $\theta^\star = \Phi^\dagger Y$. Then,

$$\begin{aligned}
\frac{1}{2}\|\Phi\theta - Y\|^2 &= \frac{1}{2}\|\Phi(\theta - \theta^\star) + \Phi\theta^\star - Y\|^2 \\
&= \frac{1}{2}\|\Phi(\theta - \theta^\star)\|^2 + (\theta - \theta^\star)^\intercal \underbrace{\Phi^\intercal(\Phi\Phi^\dagger - I)}_{=0} Y + \underbrace{\frac{1}{2}\|\Phi\theta^\star - Y\|^2}_{\stackrel{\text{def}}{=} c} \\
&= \frac{1}{2}(\theta - \theta^\star)^\intercal \underbrace{\Phi^\intercal\Phi}_{\stackrel{\text{def}}{=} H}(\theta - \theta^\star) + c.
\end{aligned}$$

Note that $H \in \mathbb{R}^{d \times d}$ is symmetric positive semidefinite.

Convex quadratic optimization

More generally, consider the optimization problem

$$\underset{\theta \in \mathbb{R}^d}{\text{minimize}} \quad F(\theta) = \tfrac{1}{2}\theta^\intercal H \theta + b^\intercal \theta + c,$$

where $H \in \mathbb{R}^{d \times d}$ is symmetric positive semidefinite, $b \in \mathbb{R}^d$, and $c \in \mathbb{R}$. Then

$$\nabla F(\theta) = H\theta + b.$$

Note that if $\theta^\intercal H \theta + b^\intercal \theta + c$ had an asymmetric $H \in \mathbb{R}^{d \times d}$, then the function is equal to

$$\frac{1}{2}\theta^\intercal (H + H^\intercal)\theta + b^\intercal \theta + c.$$

So there is no loss of generality in assuming $H \in \mathbb{R}^{d \times d}$ is symmetric. This loss function is convex if and only if $H \succeq 0$. (To be proved in homework.)

Convex quadratic optimization in standard form

For

$$F(\theta) = \frac{1}{2}\theta^\intercal H\theta + b^\intercal\theta + c,$$

there exists some $\theta^\star \in \mathbb{R}^d$ and $c' \in \mathbb{R}$ such that

$$F(\theta) = \frac{1}{2}(\theta - \theta^\star)^\intercal H(\theta - \theta^\star) + c'.$$

(To be proved in homework.) Of course, the c' is irrelevant in the optimization.

Therefore, W.L.O.G., consider

$$\underset{\theta\in\mathbb{R}^d}{\text{minimize}} \quad F(\theta) = \tfrac{1}{2}(\theta - \theta^\star)^\intercal H(\theta - \theta^\star),$$

where $H \in \mathbb{R}^{d\times d}$ is symmetric positive semidefinite. Then,

$$\nabla F(\theta) = H(\theta - \theta^\star).$$

Of course, we won't have access to $\theta^\star$, and the actual code for computing the gradient will be something like $\nabla F(\theta) = H\theta + b$. We are simply stating the mathematical fact $\nabla F(\theta) = H\theta + b = H(\theta - \theta^\star)$.

GD on strongly convex quadratic: Convergence rate

Theorem

Consider gradient descent applied to convex quadratic optimization with constant stepsize $\alpha_k = \alpha$. Let $0 < \mu \leq L < \infty$. Assume $\mu I \preceq H \preceq LI$. Then,

$$\|\theta^k - \theta^\star\|^2 \leq \Big(\max\{|1-\alpha\mu|, |1-\alpha L|\} \Big)^{2k} \|\theta^0 - \theta^\star\|^2, \qquad \text{for } k = 0, 1, \ldots.$$

If $\alpha = 2/(\mu + L)$, then,

$$\|\theta^k - \theta^\star\|^2 \leq \left(1 - \frac{2}{\kappa + 1}\right)^{2k} \|\theta^0 - \theta^\star\|^2 \leq \exp\left(-\frac{4k}{\kappa + 1}\right) \|\theta^0 - \theta^\star\|^2,$$

where $\kappa = L/\mu$, for $k = 0, 1, \ldots$.

(When $\kappa = \infty$, the bound merely guarantees $\|\theta^k - \theta^\star\|^2 \leq \|\theta^0 - \theta^\star\|$.)

GD on strongly convex quadratic: Convergence rate

Proof. Note

$$\begin{aligned}\theta^k - \theta^\star &= \theta^{k-1} - \alpha\nabla F(\theta^{k-1}) - \theta^\star \\ &= (I - \alpha H)(\theta^{k-1} - \theta^\star) = (I - \alpha H)^k(\theta^0 - \theta^\star),\end{aligned}$$

and, with $\Lambda(H)$ being the set of eigenvalues of H,

$$\min_{\lambda\in[\mu,L]}\{1-\alpha\lambda\}I \preceq \min_{\lambda\in\Lambda(H)}\{1-\alpha\lambda\}I \preceq I - \alpha H \preceq \max_{\lambda\in\Lambda(H)}\{1-\alpha\lambda\}I \preceq \max_{\lambda\in[\mu,L]}\{1-\alpha\lambda\}I.$$

This implies

$$(I - \alpha H)^2 \preceq \Big(\max_{\lambda\in[\mu,L]} |1-\alpha\lambda|\Big)^2 I.$$

Therefore,

$$\begin{aligned}\|\theta^k - \theta^\star\|^2 &= (\theta^0 - \theta^\star)^\intercal (I - \alpha H)^{2k}(\theta^0 - \theta^\star) \\ &\le \Big(\max_{\lambda\in[\mu,L]} |1-\alpha\lambda|\Big)^{2k} \|\theta^0 - \theta^\star\|^2 \\ &\le \Big(\max\{|1-\alpha\mu|, |1-\alpha L|\}\Big)^{2k} \|\theta^0 - \theta^\star\|.\end{aligned}$$

□

GD on strongly convex quadratic: Iteration complexity

Corollary

Consider gradient descent applied to convex quadratic optimization such that $\mu I \preceq H \preceq LI$. *Consider a constant stepsize* $\alpha_k = 2/(\mu + L)$. *If*

$$K \geq \frac{\kappa + 1}{4}\big(\log(1/\varepsilon) + 2\log\|\theta^0 - \theta^\star\|\big),$$

then,

$$\|\theta^K - \theta^\star\|^2 \leq \varepsilon.$$

In optimization and numerical linear algebra, $\kappa = L/\mu \geq 1$ is called the *condition number* of the problem, and it characterizes the difficulty of the problem. In algorithm design, we want efficiency for harder problem instances since easy problems are easy anyway. Therefore, we are primarily interested in the algorithm's performance when $\kappa \gg 1$.

The iteration complexity is $\mathcal{O}(\kappa)$, when all other dependences are ignored. We will soon see that $\mathcal{O}(\sqrt{\kappa})$ can be obtained via "acceleration" and that this complexity is optimal.

GD on cvx quadratics: Function values

Theorem

Consider gradient descent applied to convex quadratic optimization with constant stepsize $\alpha_k = 1/L$. Assume $\mu I \preceq H \preceq LI$. Then,

$$F(\theta^k) - F(\theta^\star) \leq \left(1 - \frac{1}{\kappa}\right)^{2k} (F(\theta^0) - F(\theta^\star)) \leq \exp\left(-\frac{2k}{\kappa}\right)(F(\theta^0) - F(\theta^\star))$$

for $k = 0, 1, \ldots$.

Proof. We first note that

$$F(\theta^k) - F(\theta^\star) = \frac{1}{2}(\theta^0 - \theta^\star)^\intercal (I - \alpha H)^{2k} H (\theta^0 - \theta^\star)$$

and

$$F(\theta^0) - F(\theta^\star) = \frac{1}{2}(\theta^0 - \theta^\star)^\intercal H (\theta^0 - \theta^\star).$$

We conclude the statement with the same line of argument as before. □

(When $\kappa = \infty$, bound merely guarantees $F(\theta^k) - F(\theta^\star) \leq F(\theta^0) - F(\theta^\star)$.)

GD on cvx quadratics: Sublinear convergence

Theorem

Consider gradient descent applied to convex quadratic optimization with constant stepsize $\alpha_k = 1/L$. Assume $0 \preceq H \preceq LI$. Then,

$$F(\theta^k) - F(\theta^\star) \leq \frac{L}{8k}\|\theta^0 - \theta^\star\|^2, \qquad \text{for } k = 0, 1, \ldots.$$

Proof. Again, note that

$$F(\theta^k) - F(\theta^\star) = \frac{1}{2}(\theta^0 - \theta^\star)^\intercal (I - \alpha H)^{2k} H (\theta^0 - \theta^\star).$$

By a similar argument as before,

$$(I - \alpha H)^{2k} H \preceq \Big(\max_{\lambda \in [0,L]} \big|\lambda(1 - \alpha\lambda)^{2k}\big| \Big) I.$$

For $\lambda \geq 0$ and $0 \leq \alpha \leq 1/L$,

$$\begin{aligned} \big|\lambda(1 - \alpha\lambda)^{2k}\big| &\leq \lambda \exp(-2k\alpha\lambda) = \frac{1}{2k\alpha} 2k\alpha\lambda \exp(-2k\alpha\lambda) \\ &\leq \frac{1}{2k\alpha} \sup_{\tau \geq 0}\{\tau \exp(-\tau)\} = \frac{1}{2ek\alpha} \leq \frac{1}{4k\alpha}. \end{aligned}$$

□

GD on cvx quadratics: Iteration complexity

Corollary

Consider gradient descent applied to convex quadratic optimization with constant stepsize $\alpha_k = 1/L$. *Assume* $0 \preceq H \preceq LI$. *If*

$$K \geq \frac{L}{8\varepsilon}\|\theta^0 - \theta^\star\|^2,$$

then

$$F(\theta^k) - F(\theta^\star) \leq \varepsilon.$$

We attain a convergence rate of $\mathcal{O}(1/k)$ and iteration complexity $\mathcal{O}(1/\varepsilon)$. We will soon see that accelerated methods achieve $\mathcal{O}(1/k^2)$ and $\mathcal{O}(1/\sqrt{\varepsilon})$ and that these are optimal.

Linear vs. sublinear convergence

In optimization, a rate of the form

$$\text{performance measure} \leq \exp(-k/C),$$

where $C > 0$, is (confusingly) referred to be both an exponential rate and a linear rate.[1]

In contrast, a rate of the form

$$\text{performance measure} \leq \frac{C}{k^\gamma},$$

where $\gamma > 0$, is said to be a sublinear rate.

[1] This is not a Q-linear rate but rather an R-linear rate. We will not worry about this distinction.

Why gradient methods?

When F is quadratic, one can use direct methods (linear algebraic) to solve the linear system $\nabla F(\theta) = 0$. Why consider gradient descent?

- When $\theta \in \mathbb{R}^d$, direct solves of $\nabla F(\theta) = 0$ requires $\mathcal{O}(d^3)$ cost, whereas $\theta^{k+1} = \theta^k - \alpha_k \nabla F(\theta^k)$ requires $\mathcal{O}(d^2)$ cost or less. (Cost per iteration is dominated by the cost of evaluating ∇F.) If GD converges to acceptable accuracy in fewer than $\mathcal{O}(d)$ iterations, GD is worthwhile.
- The guarantees of convex quadratic optimization serve as a conceptual baseline later when we try to get the same types of guarantees for convex non-quadratic optimization.

Outline

Convex non-quadratic optimization

In convex optimization, we solve

$$\underset{\theta \in \mathbb{R}^d}{\text{minimize}} \quad F(\theta),$$

where $F \colon \mathbb{R}^d \to \mathbb{R}$ is a convex function that is not necessarily a quadratic.

Since F is non-quadratic, we can no longer use the linear algebraic tools. Nevertheless, can we get the same rates as in the quadratic case?

We start by assuming F is differentiable and we later consider the case where F is non-differentiable.

GD on smooth strongly convex F: Convergence rate

We first analyze the GD with a contraction argument.

Theorem

Let $0 < \mu \le L < \infty$. Let $F\colon \mathbb{R}^d \to \mathbb{R}$ be L-smooth and μ-strongly convex. Consider GD with $\alpha_k = 1/L$. Then, for $k = 0, 1, \ldots,$

$$\|\theta^k - \theta^\star\|^2 \le (1 - 1/\kappa)^k \|\theta^0 - \theta^\star\|^2 \le \exp(-k/\kappa)\|\theta^0 - \theta^\star\|^2.$$

Proof.

$$\begin{aligned}
\|\theta^{k+1} - \theta^\star\|^2 &= \|\theta^k - \theta^\star\|^2 - 2\alpha_k \langle \nabla F(\theta^k), \theta^k - \theta^\star\rangle + \alpha_k^2 \|\nabla F(\theta^k)\|^2 \\
&\le \|\theta^k - \theta^\star\|^2 - \alpha_k(2 - \alpha_k L)\langle \nabla F(\theta^k), \theta^k - \theta^\star\rangle \\
&\le (1 - \mu\alpha_k(2 - \alpha_k L))\|\theta^k - \theta^\star\|^2 = (1 - 1/\kappa)\|\theta^k - \theta^\star\|^2,
\end{aligned}$$

where the first and second inequalities follows from

$$\langle \nabla F(\theta) - \nabla F(\eta), \theta - \eta\rangle \ge \frac{1}{L}\|\nabla F(\theta) - \nabla F(\eta)\|^2, \qquad \forall\, \theta, \eta \in \mathbb{R}^d$$

$$\langle \nabla F(\theta) - \nabla F(\eta), \theta - \eta\rangle \ge \mu\|\theta - \eta\|^2, \qquad \forall\, \theta, \eta \in \mathbb{R}^d$$

by L-smoothness and μ-strong convexity, respectively. □

GD on smooth strongly convex F: Iteration complexity

Corollary

Let $F\colon \mathbb{R}^d \to \mathbb{R}$ be L-smooth and μ-strongly convex. Consider gradient descent with constant stepsize $\alpha_k = 1/L$. Then, if

$$K \geq \kappa \log(1/\varepsilon) + 2 \log \|\theta^0 - \theta^\star\| = \mathcal{O}(\kappa \log(1/\varepsilon)),$$

then

$$\|\theta^k - \theta^\star\|^2 \leq \varepsilon.$$

Polyak–Łojasiewicz inequality

Lemma

Let $F\colon \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex, differentiable, and with a minimizer $\theta^\star$. Then,

$$\|\nabla F(\theta)\|^2 \geq 2\mu(F(\theta) - F(\theta^\star)), \qquad \forall\, \theta \in \mathbb{R}^d.$$

Proof. By μ-s.c.,

$$\begin{aligned} F(\eta) &\geq F(\theta) + \langle \nabla F(\theta), \eta - \theta \rangle + \frac{\mu}{2}\|\eta - \theta\|^2 \\ &\geq \inf_{\eta \in \mathbb{R}^d} \left\{ F(\theta) + \langle \nabla F(\theta), \eta - \theta \rangle + \frac{\mu}{2}\|\eta - \theta\|^2 \right\} = F(\theta) - \frac{1}{2\mu}\|\nabla F(\theta)\|^2. \end{aligned}$$

(Infimum is attained at $\eta = \theta - \frac{1}{\mu}\nabla F(\theta)$.) Plugging $\eta = \theta^\star$ into the LHS, we arrive at the conclusion. □

This is called the Polyak–Łojasiewicz (PL) inequality. Strong convexity implies PL. However, the converse is not true.

GD on smooth strongly convex F: Convergence rate

Using the Polyak–Łojasiewicz inequality, we can obtain a rate on F.

Theorem

Let $F\colon \mathbb{R}^d \to \mathbb{R}$ be L-smooth, convex, and μ-P.Ł. Consider gradient descent with constant stepsize $\alpha_k = 1/L$. Then, for $k = 0, 1, \ldots,$

$$F(\theta^k) - F(\theta^\star) \leq (1 - 1/\kappa)^k (F(\theta^0) - F(\theta^\star)) \leq \exp(-k/\kappa)(F(\theta^0) - F(\theta^\star)).$$

Proof.

$$\begin{aligned}
F(\theta^{k+1}) - F(\theta^\star) &= F(\theta^k - \alpha \nabla F(\theta^k)) - F(\theta^\star) \\
&\leq F(\theta^k) - \alpha \langle \nabla F(\theta^k), \nabla F(\theta^k)) \rangle + \frac{\alpha^2 L}{2} \|\nabla F(\theta^k))\|^2 - F(\theta^\star) \\
&= F(\theta^k) - F(\theta^\star) - \frac{1}{2L} \|\nabla F(\theta^k))\|^2 \\
&\leq (1 - 1/\kappa)(F(\theta^k) - F(\theta^\star)),
\end{aligned}$$

where the first inequality follow from L-smoothness and the third follows from PL. □

Specifically, the following inequality was used in the previous proof:

$$\begin{aligned} F(\theta^{k+1}) &\leq F(\theta^k) + \langle \nabla F(\theta^k), \theta^{k+1} - \theta^k \rangle + \frac{L}{2} \|\theta^{k+1} - \theta^k\|^2 \\ &= F(\theta^k) - \frac{1}{2L} \|\nabla F(\theta^k)\|^2 \end{aligned}$$

Complexity lower bound

Later, we will establish the following complexity lower bounds

Theorem

Consider first-order algorithms minimizing $F\colon \mathbb{R}^d \to \mathbb{R}$ *by only accessing* $(F(\theta^k), \nabla F(\theta^k))$ *for* $k = 0, \dots, K-1$*, where* θ^0 *is given and* $\theta^1, \dots, \theta^{K-1}$ *arechosen by the algorithm. Denote the output of the algorithm as* θ^K*. Then, there are constants* $C_1 > 0$ *and* $C_2 > 0$ *such that the following hold: for any* K *and sufficiently large* d*, there is an* L*-smooth and* μ*-strongly convex* F *such that*

$$F(\theta^K) - F(\theta^\star) \geq C_1 \exp(-C_2 K/\sqrt{\kappa}) \|\theta^0 - \theta^\star\|^2.$$

We will discuss the precise conditions of this lower bound and prove it later in this course.x For now, understand that the $\mathcal{O}(\kappa)$ iteration complexity of GD is suboptimal, and it can be accelerated.

Complexity lower bound

Later, we will establish the following complexity lower bounds

Theorem

Consider first-order algorithms minimizing $F\colon \mathbb{R}^d \to \mathbb{R}$ by only accessing $(F(\theta^k), \nabla F(\theta^k))$ for $k = 0, \dots, K-1$, where θ^0 is given and $\theta^1, \dots, \theta^{K-1}$ is chosen by the algorithm. Denote the output of the algorithm as θ^K. Then, there is a constant $C > 0$ such that the following hold: for any K and sufficiently large d, there is an L-smooth convex F such that

$$F(\theta^K) - F(\theta^\star) \geq \frac{C}{k^2} \|\theta^0 - \theta^\star\|^2.$$

Again, note that the $\mathcal{O}(1/k)$ rate of GD is suboptimal and that it can be accelerated.

GD on smooth convex F: Convergence rate

Next, we establish a sublinear rate on gradient descent.

Theorem

Let $F\colon \mathbb{R}^d \to \mathbb{R}$ be L-smooth convex. Assume F has a minimizer $\theta^\star$. Consider gradient descent with constant stepsize $\alpha_k = 1/L$. Then, for $k = 1, 2, \ldots,$

$$F(\theta^k) - F(\theta^\star) \leq \frac{L}{2k}\|\theta^0 - \theta^\star\|^2.$$

Before we get to the direct analysis of GD, let us do a warm-up exercise by considering a continuous-time model of gradient descent. Since

$$\theta^{k+1} = \theta^k - \alpha\nabla F(\theta^k) \quad \Rightarrow \quad \frac{\theta^{k+1} - \theta^k}{\alpha} = -\nabla F(\theta^k),$$

for small enough $\alpha > 0$, we can approximate the discrete-time dynamics with the continuous-time gradient flow

$$\dot{\theta}(t) = -\nabla F(\theta(t)).$$

Continuous-time analysis of GD

Consider the continuous-time dynamics

$$\dot{\theta}(t) = -\nabla F(\theta(t))$$

Define the energy function (not an obvious choice)

$$\mathcal{E}(t) = t(F(\theta) - F(\theta^\star)) + \frac{1}{2}\|\theta - \theta^\star\|^2.$$

Then, we can show $\mathcal{E}(t)$ is dissipative:

$$\begin{aligned}\frac{d}{dt}\mathcal{E}(t) &= F(\theta) - F(\theta^\star) + t\langle \nabla F(\theta), \dot{\theta}\rangle + \langle \theta - \theta^\star, \dot{\theta}\rangle \\ &= \underbrace{F(\theta) - F(\theta^\star) + \langle \theta^\star - \theta, \nabla F(\theta)\rangle}_{\leq 0 \text{ by convexity}} - t\|\nabla F(\theta)\|^2 \leq 0.\end{aligned}$$

Therefore,

$$t(F(\theta) - F(\theta^\star)) \leq \mathcal{E}(t) \leq \mathcal{E}(0) \leq \frac{1}{2}\|\theta(0) - \theta^\star\|^2$$

and we conclude

$$F(\theta) - F(\theta^\star) \leq \frac{1}{2t}\|\theta(0) - \theta^\star\|^2.$$

Discrete-time analysis of GD

Proof. Recall

$$\theta^{k+1} = \theta^k - \alpha \nabla F(\theta^k).$$

Define the energy function

$$\mathcal{E}_k = k(F(\theta^k) - F(\theta^\star)) + \frac{L}{2}\|\theta^k - \theta^\star\|^2.$$

Then, we show $\{\mathcal{E}_k\}_{k=0}^{\infty}$ is dissipative

$$\begin{aligned}
\mathcal{E}_{k+1} - \mathcal{E}_k &= (k+1)(F(\theta^{k+1}) - F(\theta^\star)) - k(F(\theta^k) - F(\theta^\star)) \\
&\qquad - \alpha L\langle F(\theta^k), \theta^k - \theta^\star\rangle + \alpha^2 L\|F(\theta^k)\|^2 \\
&\leq F(\theta^k) - F(\theta^\star) - \frac{k+1}{2L}\|\nabla F(\theta^k)\|^2 - \langle F(\theta^k), \theta^k - \theta^\star\rangle + \frac{1}{L}\|\nabla F(\theta^k)\|^2 \\
&\leq -\frac{1}{2L}\|\nabla F(\theta^k)\|^2 - \frac{k+1}{2L}\|\nabla F(\theta^k)\|^2 + \frac{1}{L}\|\nabla F(\theta^k)\|^2 \\
&= -\frac{k}{2L}\|\nabla F(\theta^k)\|^2 \leq 0,
\end{aligned}$$

where the first and second inequalities follow from L-smoothness.
The conclusion follows from

$$k(F(\theta^k) - F(\theta^\star)) \leq \mathcal{E}_k \leq \mathcal{E}_0 = \frac{L}{2}\|\theta^0 - \theta^\star\|^2.$$

□

Discrete-time analysis of GD

Specifically, the following inequalities were used in the previous proof:

$$\begin{aligned} F(\theta^{k+1}) &\leq F(\theta^k) + \langle \nabla F(\theta^k), \theta^{k+1} - \theta^k \rangle + \frac{L}{2}\|\theta^{k+1} - \theta^k\|^2 \\ &= F(\theta^k) - \frac{1}{2L}\|\nabla F(\theta^k)\|^2 \end{aligned}$$

$$F(\theta^k) - F(\theta^\star) - \langle \nabla F(\theta^k), \theta^k - \theta^\star \rangle \leq -\frac{1}{2L}\|\nabla F(\theta)\|^2.$$

Subgradient descent

Consider the optimization problem

$$\underset{\theta \in \mathbb{R}^d}{\text{minimize}} \quad F(\theta)$$

where $F : \mathbb{R}^d \to \mathbb{R}$ is convex but not differentiable.

The *subgradient descent* method is

$$g^k \in \partial F(\theta^k)$$
$$\theta^{k+1} = \theta^k - \alpha_k g^k$$

where $\theta^0 \in \mathbb{R}^d$ is a starting point and $\alpha_0, \alpha_1, \ldots \in \mathbb{R}$ is a positive sequence of stepsizes. With $g^k \in \partial F(\theta^k)$, we assume that we are given a subgradient. (We cannot choose a particular subgradient.)

(Some say the name subgradient "descent" is a misnomer since there is no guarantee that the function value $F(\theta^k)$ is monotonically decreasing.)

Lipschitz continuity $\Leftrightarrow$ bounded subgradients

Lemma

Let $F\colon \mathbb{R}^d \to \mathbb{R}$ be convex. Then, F is G-Lipschitz continuous, i.e.,

$$|F(x) - F(y)| \leq G\|x - y\|_2, \qquad \forall\, x, y \in \mathbb{R}^d$$

if and only if

$$\|\partial F(x)\|_2 \leq G, \qquad \forall\, x \in \mathbb{R}^d.$$

To clarify, L-smoothness concerns Lipschitz continuity of ∇F. Here, we are concerned with Lipschitz continuity of F.

To clarify, $\|\partial F(x)\|_2 \leq G$ means $\|g\|_2 \leq G$ for all $g \in \partial F(x)$.

If F is differentiable, then $\partial F = \nabla F$, and this lemma follows from standard calculus arguments. The proof of this lemma is beyond the scope of this course.

Convergence rate of subgradient descent

Theorem

Let $F\colon \mathbb{R}^d \to \mathbb{R}$ be a G-Lipschitz continuous convex function. Assume F has a minimizer $\theta^\star$. Let $\theta^0 \in \mathbb{R}^d$ be a starting point and write $R = \|\theta^0 - \theta^\star\|_2$. Let $K > 0$ be the total iteration count. Then, subgradient descent with the constant stepsize

$$\alpha_k = \alpha = \frac{R}{G\sqrt{K+1}}$$

exhibits the rate

$$\min_{0 \le k \le K} F(\theta^k) - f(\theta^\star) \le \frac{GR}{\sqrt{K+1}}$$

and

$$f(\bar{\theta}^K) - f(\theta^\star) \le \frac{GR}{\sqrt{K+1}},$$

where

$$\bar{\theta}^K = \frac{1}{K+1} \sum_{k=0}^{K} \theta^k.$$

Proof. For $k = 0, 1, 2, \ldots,$

$$\begin{aligned}\|\theta^{k+1} - \theta^\star\|_2^2 &= \|\theta^k - \alpha g^k - \theta^\star\|_2^2 \\ &= \|\theta^k - \theta^\star\|_2^2 - 2\alpha\langle g^k, \theta^k - \theta^\star\rangle + \alpha^2\|g^k\|_2^2 \\ &\leq \|\theta^k - \theta^\star\|_2^2 - 2\alpha(F(\theta^k) - F(\theta^\star)) + \alpha^2 G^2.\end{aligned}$$

Therefore,

$$2\alpha(F(\theta^k) - F(\theta^\star)) \leq \|\theta^k - \theta^\star\|_2^2 - \|\theta^{k+1} - \theta^\star\|_2^2 + \alpha^2 G^2.$$

With a telescoping sum argument, we get

$$\begin{aligned}2\alpha\sum_{k=0}^{K}(F(\theta^k) - F(\theta^\star)) &\leq \|\theta^0 - \theta^\star\|_2^2 - \|\theta^{K+1} - \theta^\star\|_2^2 + \sum_{k=0}^{K}\alpha^2 G^2 \\ &\leq R^2 + (K+1)\alpha^2 G^2,\end{aligned}$$

and

$$\frac{1}{K+1}\sum_{k=0}^{K} F(\theta^k) - F(\theta^\star) \leq \frac{R^2 + \alpha^2 G^2 (K+1)}{2\alpha(K+1)} = \frac{GR}{\sqrt{K+1}}.$$

Therefore,

$$\min_{0\leq k\leq K} F(\theta^k) - F(\theta^\star) = \frac{1}{K+1}\sum_{k=0}^{K} \min_{0\leq k\leq K} F(\theta^k) - F(\theta^\star)$$
$$\leq \frac{1}{K+1}\sum_{k=0}^{K} F(\theta^k) - F(\theta^\star) \leq \frac{GR}{\sqrt{K+1}}.$$

Likewise, using Jensen's inequality, we conclude

$$F(\bar{\theta}^K) - F(\theta^\star) \leq \frac{1}{K+1}\sum_{k=0}^{K} F(\theta^k) - F(\theta^\star) \leq \frac{GR}{\sqrt{K+1}}.$$

□

Complexity lower bound

Subgradient descent exhibits a $\mathcal{O}(1/\sqrt{K})$ rate, which is slower than gradient descent on L-smooth convex functions. This $\mathcal{O}(1/\sqrt{K})$ rate turns out to be optimal.

Theorem

Consider first-order algorithms minimizing $F\colon \mathbb{R}^d \to \mathbb{R}$ by only accessing $(F(\theta^k), g^k)$, where $g^k \in \partial F(\theta^k)$ is a subgradient, for $k = 0, \dots, K-1$, where θ^0 is given and $\theta^1, \dots, \theta^{K-1}$ is chosen by the algorithm. Denote the output of the algorithm as θ^K. Then, there is a constant $C > 0$ such that the following hold: for any K and sufficiently large d, there is an G-Lipschitz convex F such that

$$F(\theta^K) - F(\theta^\star) \geq \frac{C}{\sqrt{K}} \|\theta^0 - \theta^\star\|^2.$$

Projected gradient descent

Consider the problem

$$\begin{array}{ll} \underset{\theta \in \mathbb{R}^d}{\text{minimize}} & F(\theta) \\ \text{subject to} & \theta \in \Theta, \end{array}$$

where $\Theta \subseteq \mathbb{R}^d$ is a nonempty closed convex set and $F \colon \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable.

The method *projected* gradient method is

$$\theta^{k+1} = \mathrm{Proj}_\Theta(\theta^k - \alpha_k \nabla F(\theta^k))$$

for $k = 0, 1, \ldots$, where the projection operator is defined as

$$\mathrm{Proj}_\Theta(\theta_0) = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{2}\|\theta - \theta_0\|^2.$$

Proximal gradient descent

Consider the problem

$$\underset{\theta \in \mathbb{R}^d}{\text{minimize}} \quad F(\theta) + H(\theta),$$

where $F \colon \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable and $H \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is CCP.

The method *proximal* gradient method is

$$\theta^{k+1} = \mathrm{Prox}_{\alpha^k H}(\theta^k - \alpha_k \nabla F(\theta^k))$$

for $k = 0, 1, \ldots$, where $\theta^0 \in \mathbb{R}^d$, $\alpha_k > 0$ for $k = 0, 1, \ldots$ and the *proximal operator* is defined as

$$\mathrm{Prox}_{\alpha H}(\theta_0) = \underset{\theta \in \mathbb{R}^d}{\mathrm{argmin}} \left\{ \alpha H(\theta) + \frac{1}{2} \|\theta - \theta_0\|^2 \right\}.$$

Prox generalizes projection

Let $\Theta \subseteq \mathbb{R}^d$. Let $\delta_\Theta : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be defined as

$$\delta_\Theta(\theta) = \begin{cases} 0 & \text{if } \theta \in \Theta \\ \infty & \text{if } \theta \notin \Theta \end{cases}$$

We call δ_Θ the *indicator function* with respect to Θ. If $\Theta \subseteq \mathbb{R}^d$ is a nonempty closed convex set, then δ_Θ is CCP.

Using the indicator function, we can convert a constrained optimization problem into an unconstrained one:

$$\begin{array}{ll} \underset{\theta \in \mathbb{R}^d}{\text{minimize}} & F(\theta) \\ \text{subject to} & \theta \in \Theta \end{array}$$

is equivalent to

$$\underset{\theta \in \mathbb{R}^d}{\text{minimize}} \quad F(\theta) + \delta_\Theta(\theta).$$

Prox generalizes projection

If $H = \delta_\Theta$ and $\alpha > 0$, then

$$\begin{aligned}\text{Prox}_{\alpha H}(\theta_0) &= \underset{\theta \in \mathbb{R}^d}{\text{argmin}} \left\{ \alpha H(\theta) + \frac{1}{2}\|\theta - \theta_0\|^2 \right\} \\ &= \underset{\theta \in \Theta}{\text{argmin}} \left\{ \frac{1}{2}\|\theta - \theta_0\|^2 \right\} = \text{Proj}_\Theta(\theta_0).\end{aligned}$$

So, the proximal operator generalizes the projection operator, and proximal GD generalizes projected GD.

Prox is well defined

Lemma

If $H : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ *is CCP and* $\alpha > 0$*, then*

$$\mathrm{Prox}_{\alpha H}(\theta_0) = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \alpha H(\theta) + \frac{1}{2} \|\theta - \theta_0\|^2 \right\}.$$

is well defined, i.e., the argmin uniquely exists.

Proof. In Chapter 0. □

Lemma

Let $F : \mathbb{R}^d \to \mathbb{R}$ be CCP. Then

$$\theta^\star \in \operatorname{argmin} F \quad \Leftrightarrow \quad 0 \in \partial F(\theta^\star).$$

Proof. Immediate from definition. □

Lemma

Let $F : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable and $H : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be CCP. Then

$$\partial(F + H)(\theta) = \nabla F(\theta) + \partial H(\theta) = \{\nabla F(\theta) + g \,|\, g \in \partial H(\theta)\}.$$

Proof. In Chapter 0. □

Prox-grad encodes a solution as a fixed point

Lemma

If prox-GD starts at a solution, then it does not move, i.e.,
if $\theta^0 \in \operatorname{argmin}(F + H)$*, then* $\theta^1 = \theta^0$*.*

Proof. Note,

$$\theta^1 = \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \alpha H(\theta) + \frac{1}{2} \|\theta - (\theta^0 - \alpha \nabla F(\theta^0))\|^2 \right\}$$

is equivalent to

$$0 \in \alpha \partial H(\theta^1) + \theta^1 - \theta^0 + \alpha \nabla F(\theta^0).$$

If θ^0 is a solution, then $0 \in \partial H(\theta^0) + \nabla F(\theta^0)$ and

$$0 \in \alpha \partial H(\theta^1) + \theta^1 - \theta^0 + \alpha \nabla F(\theta^0).$$

holds with $\theta^1 = \theta^0$, so $\theta^1 = \theta^0$ is a minimizer of $\operatorname{argmin}_{\theta \in \mathbb{R}^d} \{\cdots\}$. Since the prox is uniquely defined, we conclude $\theta^1 = \theta^0$. □

Lemma

If prox-GD does not move, then it is at a solution, i.e.,
if $\theta^1 = \theta^0$*, then* $\theta^0 \in \operatorname{argmin}(F + H)$*.*

Proof. If $\theta^1 = \theta^0$, then $0 \in \partial H(\theta^0) + \nabla F(\theta^0)$ and θ^0 is a solution. □

Prox-GD: Convergence rate

Next, we establish a sublinear rate on proximal gradient descent.

Theorem

Let $F : \mathbb{R}^d \to \mathbb{R}$ be L-smooth convex and $H : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be CCP. Assume $F + H$ has a minimizer $\theta^\star$. Consider proximal gradient descent with constant stepsize $\alpha_k = 1/L$. Then, for $k = 1, 2, \ldots,$

$$F(\theta^k) + H(\theta^k) - F(\theta^\star) - H(\theta^\star) \leq \frac{L}{2k} \|\theta^0 - \theta^\star\|^2.$$

Note that this is the same rate as GD.

Lemma

Let $F\colon \mathbb{R}^d \to \mathbb{R}$ be L-smooth convex and $H\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ be CCP. Let $\theta^+ = \operatorname{Prox}_{\frac{1}{L}H}\big(\theta - \frac{1}{L}\nabla F(\theta)\big)$. Define $G = F + H$. Then for any $\varphi \in \mathbb{R}^d$, we have

$$G(\theta^+) + L\langle \theta - \theta^+, \varphi - \theta\rangle + \frac{L}{2}\|\theta^+ - \theta\|^2 \leq G(\varphi).$$

Corollary

With $\varphi = \theta$, we have

$$G(\theta^+) \leq G(\theta) - \frac{L}{2}\|\theta^+ - \theta\|^2.$$

Corollary

With $\varphi = \theta^\star \in \operatorname{argmin} G$, we have

$$G(\theta^+) - G(\theta^\star) + L\langle \theta - \theta^+, \theta^\star - \theta\rangle \leq -\frac{L}{2}\|\theta^+ - \theta\|^2.$$

Proof. Since F is L-smooth convex, we have

$$F(\theta^+) + H(\theta^+) \leq \langle \theta^+ - \theta, \nabla F(\theta) \rangle - \frac{L}{2}\|\theta^+ - \theta\|^2 + F(\theta) + H(\theta^+).$$

By convexity, we have

$$F(\theta) \leq F(\varphi) - \langle \varphi - \theta, \nabla F(\theta) \rangle.$$

By the subgradient inequality, we have

$$H(\theta^+) \leq H(\varphi) - \langle g, \varphi - \theta^+ \rangle, \quad g \in \partial H(\theta^+).$$

The optimality condition for

$$\theta^+ = \operatorname*{argmin}_{\varphi \in \mathbb{R}^d} \left\{ H(\varphi) + \tfrac{L}{2} \left\| \varphi - \left(\theta - \tfrac{1}{L} \nabla F(\theta) \right) \right\|^2 \right\}$$

implies

$$\partial H(\theta^+) + L(\theta^+ - \theta) + \nabla F(\theta) \ni 0,$$

So

$$H(\theta^+) \leq H(\varphi) + \langle L(\theta^+ - \theta) + \nabla F(\theta), \varphi - \theta^+ \rangle.$$

Combining the bounds, we conclude the stated result. □

Analysis of prox-GD

Proof. Define $G(\theta) = F(\theta) + H(\theta) - F(\theta^\star) - H(\theta^\star)$.

Define the energy function

$$\mathcal{E}^k = kG(\theta^k) + \frac{L}{2}\|\theta^k - \theta^\star\|^2.$$

If we show $\{\mathcal{E}^k\}_{k=0}^\infty$ is dissipative, then the conclusion follows from

$$kG(\theta^k) \leq \mathcal{E}^k \leq \mathcal{E}^0 = \frac{L}{2}\|\theta^0 - \theta^\star\|^2.$$

Analysis of prox-GD

$$
\begin{aligned}
&\mathcal{E}^{k+1} - \mathcal{E}^k \\
&= (k+1)G(\theta^{k+1}) - kG(\theta^k) + \frac{L}{2}\|\theta^{k+1} - \theta^\star\|^2 - \frac{L}{2}\|\theta^k - \theta^\star\|^2 \\
&= (k+1)G(\theta^{k+1}) - kG(\theta^k) + \frac{L}{2}\langle \theta^{k+1} - \theta^k, \theta^{k+1} + \theta^k - 2\theta^\star\rangle \\
&= (k+1)G(\theta^{k+1}) - kG(\theta^k) + \frac{L}{2}\|\theta^{k+1} - \theta^k\|^2 + L\langle \theta^{k+1} - \theta^k, \theta^k - \theta^\star\rangle
\end{aligned}
$$

Using

$$
G(\theta^{k+1}) \leq G(\theta^k) - \frac{L}{2}\|\theta^{k+1} - \theta^k\|^2
$$

and

$$
G(\theta^{k+1}) + L\langle \theta - \theta^{k+1}, \theta^\star - \theta^k\rangle \leq -\frac{L}{2}\|\theta^{k+1} - \theta^k\|^2,
$$

we conclude

$$
\mathcal{E}^{k+1} - \mathcal{E}^k \leq -\frac{Lk}{2}\|\theta^{k+1} - \theta^k\|^2 \leq 0.
$$

□

Accelerated gradient method

Consider

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad f(x),$$

where f is L-smooth convex. The method

$$x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$$
$$y^{k+1} = x^{k+1} + \frac{k-1}{k+2}(x^{k+1} - x^k)$$

for $k = 0, 1, \ldots$, where $x^0 = y^0 \in \mathbb{R}^d$, is Nesterov's accelerated gradient method (AGM).

Theorem

Assume the convex, L-smooth function f has a minimizer $x^\star$. Then AGM converges with the rate

$$f(x^k) - f(x^\star) \leq \frac{2L\|x^0 - x^\star\|^2}{k^2}, \qquad \text{for } k = 1, 2, \ldots.$$

Convergence analysis of AGM

Lemma

Equivalent form of AGM:

$$x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$$
$$z^{k+1} = z^k - \frac{k+1}{2L}\nabla f(y^k)$$
$$y^{k+1} = \left(1 - \frac{2}{k+2}\right) x^{k+1} + \frac{2}{k+2} z^{k+1}$$

for $k = 0, 1, \ldots,$ *where* $x^0 = y^0 = z^0 \in \mathbb{R}^d$.

Proof. Follows from induction. □

Preliminary observations. Define

$$\theta_k = \frac{k+1}{2}$$

for $k = -1, 0, 1, \ldots$. It is straightforward to verify

$$\theta_k^2 - \theta_k \leq \theta_{k-1}^2 \tag{1}$$

for $k = 0, 1, \ldots$. We will use the inequalities

$$f(x^{k+1}) - f(y^k) + \frac{1}{2L}\|\nabla f(y^k)\|^2 \leq 0 \tag{2}$$

$$f(y^k) - f(x^k) \leq \langle \nabla f(y^k), y^k - x^k \rangle \tag{3}$$

$$f(y^k) - f(x^\star) \leq \langle \nabla f(y^k), y^k - x^\star \rangle. \tag{4}$$

The first, (2), follows from L-smoothness, which implies $f(x) - \frac{L}{2}\|x - y^k\|^2$ is concave as a function of x, which in turn implies

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

We plug in $x = x^{k+1} = y^k - \frac{1}{L}\nabla f(y^k)$ to get (2). The second and third inequalities, (3) and (4), follow from convexity of f.

Convergence analysis of AGM

Define

$$\mathcal{E}^k = \theta_{k-1}^2 \left(f(x^k) - f(x^\star)\right) + \frac{L}{2}\|z^k - x^\star\|^2.$$

If we establish $\mathcal{E}^k \leq \mathcal{E}^{k-1} \leq \cdots \leq \mathcal{E}^0$, then $\mathcal{E}^k \leq \mathcal{E}^0$ implies

$$\theta_{k-1}^2(f(x^k) - f(x^\star)) \leq \mathcal{E}^k \leq \mathcal{E}^0 = \frac{L}{2}\|z^0 - x^\star\|^2.$$

$$
\begin{aligned}
&\mathcal{E}^{k+1} - \mathcal{E}^k \\
&= \theta_k^2 \left(f(x^{k+1}) - f(x^\star) + \frac{1}{2L}\|\nabla f(y^k)\|^2 \right) - \theta_{k-1}^2 (f(x^k) - f(x^\star)) \\
&\qquad - \theta_k \langle \nabla f(y^k), z^k - x^\star \rangle \\
&\overset{(2)}{\leq} \theta_k^2 \Big(f(y^k) - f(x^\star) \Big) - \theta_{k-1}^2 (f(x^k) - f(x^\star)) - \theta_k \langle \nabla f(y^k), z^k - x^\star \rangle \\
&= (\theta_k^2 - \theta_k)(f(y^k) - f(x^k)) + \theta_k (f(y^k) - f(x^k)) + (\theta_k^2 - \theta_{k-1}^2)(f(x^k) - f(x^\star)) \\
&\qquad - \theta_k \langle \nabla f(y^k), z^k - x^\star \rangle \\
&\overset{(1)}{\leq} (\theta_k^2 - \theta_k)(f(y^k) - f(x^k)) + \theta_k (f(y^k) - f(x^\star)) - \theta_k \langle \nabla f(y^k), z^k - x^\star \rangle \\
&\overset{(3),(4)}{\leq} (\theta_k^2 - \theta_k) \langle \nabla f(y^k), y^k - x^k \rangle + \theta_k \langle \nabla f(y^k), y^k - x^\star \rangle - \theta_k \langle \nabla f(y^k), z^k - x^\star \rangle \\
&= \theta_k \langle \nabla f(y^k), (1-\theta_k) x^k + \theta_k y^k - z^k \rangle \overset{\text{def. of } z^k}{=} 0,
\end{aligned}
$$

where the first equality follows from

$$
\frac{L}{2} \left\| z^k - x^\star - \frac{\theta_k}{L} \nabla f(y^k) \right\|^2 - \frac{L}{2} \|z^k - x^\star\|^2 = -\theta_k \langle \nabla f(y^k), z^k - x^\star \rangle + \frac{\theta_k^2}{2L} \|\nabla f(y^k)\|^2.
$$

□

Accelerated proximal gradient (FISTA)

Consider the problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad f(x)+g(x),$$

where f is differentiable convex and g is CCP. The method

$$x^{k+1} = \mathbf{prox}_{\frac{1}{L}g}\left(y^k - \tfrac{1}{L}\nabla f(y^k)\right)$$
$$y^{k+1} = x^{k+1} + \frac{k-1}{k+2}(x^{k+1}-x^k)$$

for $k=0,1,\ldots$, where $x^0=y^0\in\mathbb{R}^d$, is called the accelerated proximal gradient method or fast iterative shrinkage-thresholding algorithm (FISTA).

Convergence analysis of FISTA

Theorem

Assume f is L-smooth convex and g is CCP. Assume $f+g$ has a minimizer $x^\star$. Then FISTA converges with the rate

$$f(x^k)+g(x^k)-f(x^\star)-g(x^\star) \leq \frac{2L\|x^0-x^\star\|^2}{k^2}, \qquad \text{for } k=1,2,\ldots.$$

Proof. Define $\theta_k = \frac{k+1}{2}$ for $k=-1,0,1,\ldots$. Let

$$\mathcal{E}^k = \theta_{k-1}^2\left(f(x^k)+g(x^k)-f(x^\star)-g(x^\star)\right)+\frac{L}{2}\|z^k-x^\star\|^2.$$

If we establish $\mathcal{E}^k \leq \mathcal{E}^{k-1} \leq \cdots \leq \mathcal{E}^0$, then $\mathcal{E}^k \leq \mathcal{E}^0$ implies

$$\theta_{k-1}^2(f(x^k)-f(x^\star)) \leq \mathcal{E}^k \leq \mathcal{E}^0 = \frac{L}{2}\|x^0-x^\star\|^2.$$

Convergence analysis of FISTA

Lemma

Equivalent form of FISTA:

$$
\begin{aligned}
x^{k+1} &= \mathbf{prox}_{\frac{1}{L}g}\left(y^k - \tfrac{1}{L}\nabla f(y^k)\right) \\
z^{k+1} &= z^k + \theta_k(x^{k+1} - y^k) \\
y^{k+1} &= \left(1 - \frac{1}{\theta_{k+1}}\right) x^{k+1} + \frac{1}{\theta_{k+1}} z^{k+1}
\end{aligned}
$$

for $k = 0, 1, \ldots,$ *where* $x^0 = y^0 = z^0 \in \mathbb{R}^d$.

Proof. Follows from induction. □

Convergence analysis of FISTA

Preliminary observations. Define

$$\theta_k = \frac{k+1}{2}$$

for $k = -1, 0, 1, \ldots$. It is straightforward to verify

$$\theta_k^2 - \theta_k \leq \theta_{k-1}^2 \tag{1}$$

for $k = 0, 1, \ldots$.

Let $F = f + g$. We will use the inequalities

$$F(x^{k+1}) - F(x^\star) \leq L\langle x^\star - y^k, x^{k+1} - y^k\rangle - \frac{L}{2}\|x^{k+1} - y^k\|^2 \tag{2}$$

$$F(x^{k+1}) - F(x^k) \leq L\langle x^k - y^k, x^{k+1} - y^k\rangle - \frac{L}{2}\|x^{k+1} - y^k\|^2 \tag{3}$$

which follows from plugging in $\varphi = x^\star$ and $\varphi = x^k$ into the lemma we proved for the analysis of prox-grad.

$$
\begin{aligned}
&\mathcal{E}^{k+1} - \mathcal{E}^k \\
&= \theta_k^2(F(x^{k+1}) - F(x^\star)) - \theta_{k-1}^2(F(x^k) - F(x^\star)) + \frac{L}{2}\|z^{k+1} - x^\star\|^2 - \frac{L}{2}\|z^k - x^\star\|^2 \\
&\overset{\bullet}{=} \theta_k^2(F(x^{k+1}) - F(x^\star)) - \theta_{k-1}^2(F(x^k) - F(x^\star)) \\
&\qquad + L\theta_k\langle x^{k+1} - y^k, z^k - x^\star\rangle + \frac{L\theta_k^2}{2}\|x^{k+1} - y^k\|^2 \\
&\overset{(1)}{\leq} \theta_k^2(F(x^{k+1}) - F(x^\star)) - \theta_k(\theta_k - 1)(F(x^k) - F(x^\star)) \\
&\qquad + \frac{L\theta_k^2}{2}\|x^{k+1} - y^k\|^2 + L\theta_k\langle x^{k+1} - y^k, z^k - x^\star\rangle \\
&= (\theta_k^2 - \theta_k)(F(x^{k+1}) - F(x^k)) + \theta_k(F(x^{k+1}) - F(x^\star)) + \frac{L(\theta_k^2 - \theta_k)}{2}\|x^{k+1} - y^k\|^2 \\
&\qquad + \frac{L\theta_k}{2}\|x^{k+1} - y^k\|^2 + L\theta_k\langle x^{k+1} - y^k, \underbrace{\theta_k y^k - (\theta_k - 1)x^k}_{\text{def. of } z^k} - x^\star\rangle
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(3)}{\leq} L\theta_k(\theta_k-1)\langle x^k-y^k, x^{k+1}-y^k\rangle + \theta_k(F(x^{k+1})-F(x^\star)) \\
&\qquad + \frac{L\theta_k}{2}\|x^{k+1}-y^k\|^2 + L\theta_k\langle x^{k+1}-y^k, \theta_k y^k-(\theta_k-1)x^k-x^\star\rangle \\
&= \theta_k(F(x^{k+1})-F(x^\star)) + \frac{L\theta_k}{2}\|x^{k+1}-y^k\|^2 + L\theta_k\langle x^{k+1}-y^k, y^k-x^\star\rangle \\
&\overset{(2)}{\leq} 0.
\end{aligned}
$$

where $(\bullet)$ follows from

$$
\frac{L}{2}\left\|z^k-x^\star-\theta_k(x^{k+1}-y^k)\right\|^2 - \frac{L}{2}\|z^k-x^\star\|^2 = L\theta_k\langle x^{k+1}-y^k, z^k-x^\star\rangle + \frac{L\theta_k^2}{2}\|x^{k+1}-y^k\|^2.
$$

□

Strongly-convex AGM

Consider the problem

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad f(x),$$

where f is μ-strongly convex and L-smooth. The method

$$x^{k+1} = y^k - \frac{1}{L} \nabla f(y^k)$$
$$y^{k+1} = x^{k+1} + \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}(x^{k+1} - x^k)$$

for $k = 0, 1, \ldots$, where $x^0 = y^0 \in \mathbb{R}^d$, and $\kappa = L/\mu$, is called the strongly convex accelerated gradient method (SC-AGM).

Theorem

Let $0 < \mu < L < \infty$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and μ-strongly convex. Then, SC-AGM converges with the rate

$$f(x^k) - f(x^\star) \leq \frac{\mu + L}{2}\|x^0 - x^\star\|^2 e^{-k/\sqrt{\kappa}}, \qquad \text{for } k = 0, 1, \ldots.$$

Proof. Let $z^k = (1 + \sqrt{\kappa})y^k - \sqrt{\kappa}x^k$ and

$$\mathcal{E}^k = \left(1 + \frac{1}{\sqrt{\kappa} - 1}\right)^k \left(f(x^k) - f(x^\star) + \frac{\mu}{2}\|z^k - x^\star\|^2\right)$$

for $k = 0, 1, \ldots$. If we establish $\mathcal{E}^k \leq \cdots \leq \mathcal{E}^0$, then $\mathcal{E}^k \leq \mathcal{E}^0$ implies

$$\begin{aligned}\left(1 + \frac{1}{\sqrt{\kappa} - 1}\right)^k (f(x^k) - f(x^\star)) &\leq \mathcal{E}^k \leq \mathcal{E}^0 \\ = f(x^0) - f(x^\star) + \frac{\mu}{2}\|x^0 - x^\star\|^2 &\leq \frac{L}{2}\|x^0 - x^\star\|^2 + \frac{\mu}{2}\|x^0 - x^\star\|^2.\end{aligned}$$

We conclude the statement with

$$\left(1 + \frac{1}{\sqrt{\kappa} - 1}\right)^{-k} \leq \exp\left(\frac{-k}{\sqrt{\kappa} - 1}\right) \leq \exp\left(\frac{-k}{\sqrt{\kappa}}\right).$$

□

Outline

Stochastic optimization

Consider the stochastic optimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad \mathop{\mathbb{E}}_{\omega}[f(x;\omega)] = F(x),$$

where ω is a random variable. In machine learning, such problems arise in the finite-sum form

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad \mathop{\mathbb{E}}_{I\sim\text{Uniform}\{1,\dots,N\}}[f_I(x)] = \frac{1}{N}\sum_{i=1}^{N} f_i(x),$$

or

$$\underset{\theta\in\mathbb{R}^d}{\text{minimize}} \quad \frac{1}{N}\sum_{i=1}^{N} \ell(f_\theta(X_i), Y_i).$$

Stochastic (sub)gradients

Under mild conditions, we have

$$\nabla F(x) = \nabla \mathop{\mathbb{E}}_{\omega}[f(x;\omega)] = \mathop{\mathbb{E}}_{\omega}[\nabla_x f(x;\omega)].$$

Therefore, $\nabla_x f(x;\omega)$ is an unbiased estimate of $\nabla F(x)$, and we say $\nabla_x f(x;\omega)$ is a *stocahstic gradient* of F at $x \in \mathbb{R}^d$.

Let $g_\omega \in \partial f(x;\omega)$ be a random subgradient at $x \in \mathbb{R}^d$. Then,

$$\begin{aligned} F(y) = \mathop{\mathbb{E}}_{\omega}[f(y;\omega)] &\geq \mathop{\mathbb{E}}_{\omega}[f(x;\omega) + \langle g_\omega, y - x\rangle] \\ &= F(x) + \langle \mathop{\mathbb{E}}_{\omega}[g_\omega], y - x\rangle, \qquad \forall\, y \in \mathbb{R}^d \end{aligned}$$

and $\mathbb{E}_\omega[g_\omega] \in \partial F(x)$, provided that $\mathbb{E}_\omega[g_\omega]$ is well defined. In this case, we say $g_\omega \in \partial f(x;\omega)$ is a *stocahstic subgradient* of F at $x \in \mathbb{R}^d$.

Stochastic (sub)gradient descent (SGD)

Consider the algorithm stochastic (sub)gradient descent (SGD)

$$x^{k+1} = x^k - \alpha_k g^k$$

for $k = 0, 1, \ldots$, where g^k is a stochastic (sub)gradient of F at x^k.

More specifically, we assume that

$$\mathbb{E}_k[g^k] \in \partial F(x^k),$$

where

$$\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \,|\, x^0, x^1, \ldots, x^k]$$

is the conditional expectation, conditioned on the iterates up to x^k. We will also assume that the conditional variance is bounded:

$$\mathrm{Var}_k(g^k) = \mathbb{E}_k\left[\|g^k - \mathbb{E}_k[g^k]\|^2\right] \leq \sigma^2$$

Analysis of SGD

Theorem

Let $F\colon \mathbb{R}^d \to \mathbb{R}$ be a G-Lipschitz continuous convex function. Assume F has a minimizer $x^\star$. Let $x^0 \in \mathbb{R}^d$ be a starting point. Let $K > 0$ be the total iteration count. Assume the stochastic subgradient g^k satisfies

$$\mathbb{E}_k[g^k] \in \partial F(x^k), \qquad \mathrm{Var}_k(g^k) \leq \sigma^2$$

for $k = 0, 1, \ldots$. Then, SGD with the constant stepsize

$$\alpha_k = \alpha = \frac{\|x^0 - x^\star\|_2}{\sqrt{G^2 + \sigma^2}\sqrt{K+1}}$$

exhibits the rate

$$\mathbb{E}\big[f(\bar{x}^K) - f(x^\star)\big] \leq \frac{\sqrt{G^2 + \sigma^2}\|x^0 - x^\star\|_2}{\sqrt{K+1}},$$

where

$$\bar{x}^K = \frac{1}{K+1}\sum_{k=0}^{K} x^k.$$

Analysis of SGD

Proof. First,

$$\begin{aligned}\mathbb{E}_k\left[\|x^{k+1}-x^\star\|_2^2\right] &= \|x^k-x^\star\|_2^2 - 2\alpha\langle\mathbb{E}_k[g^k], x_k-x^\star\rangle + \alpha^2\mathbb{E}_k[\|g^k\|^2] \\ &\leq \|x^k-x^\star\|_2^2 - 2\alpha(F(x^k)-F(x^\star)) + \alpha^2(G^2+\sigma^2).\end{aligned}$$

We take the total expectation on both sides to get

$$\mathbb{E}\left[\|x^{k+1}-x^\star\|_2^2\right] \leq \mathbb{E}\left[\|x^k-x^\star\|_2^2\right] - 2\alpha\mathbb{E}[F(x^k)-F(x^\star)] + \alpha^2(G^2+\sigma^2).$$

By the same telescoping-sum argument as in the (non-stochastic) subgradient descent, we have

$$\mathbb{E}[F(\bar{x}^K)-F(x^\star)] \leq \frac{1}{K+1}\sum_{k=0}^{K}\mathbb{E}[F(x^k)-F(x^\star)] \leq \frac{\sqrt{G^2+\sigma^2}R}{\sqrt{K+1}}.$$

□

Projected stochastic gradient

Let $C \subseteq \mathbb{R}^d$ be a nonempty closed convex set. Consider the stochastic optimization problem

$$\begin{array}{ll} \underset{x \in \mathbb{R}^d}{\text{minimize}} & \underset{\omega}{\mathbb{E}}[f(x;\omega)] = F(x) \\ \text{subject to} & x \in C, \end{array}$$

where ω is a random variable.

Consider the algorithm *projected* stochastic (sub)gradient descent (SGD)

$$x^{k+1} = \mathrm{Proj}_C(x^k - \alpha_k g^k)$$

for $k = 0, 1, \ldots$, where g^k is a stochastic (sub)gradient of F at x^k.

Analysis of projected SGD

Theorem

Let $F\colon \mathbb{R}^d \to \mathbb{R}$ be a G-Lipschitz continuous convex function. Assume $x^\star \in \operatorname{argmin}_{x\in C} F(x)$ exists. Let $x^0 \in \mathbb{R}^d$ be a starting point. Let $K > 0$ be the total iteration count. Assume the stochastic subgradient g^k satisfies

$$\mathbb{E}_k[g^k] \in \partial F(x^k), \qquad \operatorname{Var}_k(g^k) \le \sigma^2$$

for $k = 0, 1, \ldots$. Then, projected SGD with the constant stepsize

$$\alpha_k = \alpha = \frac{\|x^0 - x^\star\|_2}{\sqrt{G^2 + \sigma^2}\sqrt{K+1}}$$

exhibits the rate

$$\mathbb{E}\big[F(\bar{x}^K) - F(x^\star)\big] \le \frac{\sqrt{G^2+\sigma^2}\|x^0 - x^\star\|_2}{\sqrt{K+1}},$$

where

$$\bar{x}^K = \frac{1}{K+1}\sum_{k=0}^{K} x^k.$$

Analysis of projected SGD

Proof. By nonexpansivity of projection,

$$\begin{aligned}\|x^{k+1} - x^\star\|_2^2 &= \|\mathrm{Proj}_C(x^k - \alpha g^k) - \mathrm{Proj}_C(x^\star)\|_2^2 \\ &\leq \|x^k - \alpha g^k - x^\star\|_2^2.\end{aligned}$$

Now the analysis proceed the same as before

$$\begin{aligned}\mathbb{E}_k\left[\|x^{k+1} - x^\star\|_2^2\right] &\leq \mathbb{E}_k\left[\|x^k - \alpha g^k - x^\star\|_2^2\right] \\ &= \|x^k - x^\star\|_2^2 - 2\alpha\langle \mathbb{E}_k[g^k], x_k - x^\star\rangle + \alpha^2 \mathbb{E}_k[\|g^k\|^2] \\ &\leq \|x^k - x^\star\|_2^2 - 2\alpha(F(x^k) - F(x^\star)) + \alpha^2(G^2 + \sigma^2).\end{aligned}$$

□

Complexity lower bound

Later we will establish a complexity lower bound showing that

$$\mathcal{O}\Big(\frac{1}{\sqrt{K}}\Big)$$

is the correct (optimal rate). This rate will remain the same under the assumption that F is L-smooth convex.

When F is strongly convex, a lower bound establishes that

$$\mathcal{O}\Big(\frac{1}{K}\Big)$$

is the correct (optimal rate). In the following, we will carry out a simple but slightly suboptimal analysis to show a rate of order

$$\mathcal{O}\Big(\frac{\log K}{K}\Big).$$

Strongly convex SGD

Consider the stochastic optimization problem

$$\begin{array}{ll} \underset{x\in\mathbb{R}^d}{\text{minimize}} & \mathop{\mathbb{E}}_{\omega}[f(x;\omega)] + \dfrac{\mu}{2}\|x\|^2 = F(x) + \dfrac{\mu}{2}\|x\|^2 \\ \text{subject to} & x \in C, \end{array}$$

where ω is a random variable. We assume F is Lipschitz continuous. ($F(x) + \frac{\mu}{2}\|x\|^2$ cannot be Lipschitz continuous.)

Consider the stochastic (sub)gradient descent (SGD)

$$x^{k+1} = x^k - \alpha_k(g^k + \mu x^k)$$

for $k = 0, 1, \ldots$, where g^k is a stochastic (sub)gradient of F at x^k.

Strongly convex SGD

Theorem

Let $F\colon \mathbb{R}^d \to \mathbb{R}$ be a G-Lipschitz continuous convex function. Let $x^\star$ be the minimizer of $F(x) + \frac{\mu}{2}\|x\|^2$. Let $x^0 \in \mathbb{R}^d$ be the starting point. Assume the stochastic subgradient g^k satisfies

$$\mathbb{E}_k[g^k] \in \partial F(x^k), \qquad \mathrm{Var}_k(g^k) \leq \sigma^2$$

for $k = 0, 1, \ldots$. Then, SGD with stepsize

$$\alpha_k = \frac{1}{\mu(k+1)}$$

exhibits the rate

$$\mathbb{E}\big[F(\bar{x}^k) - F(x^\star)\big] \leq \frac{2(G^2+\sigma^2)}{\mu}\frac{1+\log(k+1)}{k+1},$$

where

$$\bar{x}^k = \frac{1}{k+1}\sum_{s=0}^{k} x^s.$$

Proof. Let $\mathbb{E}_k[g^k] = G^k \in \partial F(x^k)$. Since $x^0 = 0$,

$$\begin{aligned} x^{k+1} &= x^k - \alpha_k(g^k + \mu x^k) \\ &= (1 - \alpha_k\mu)x^k + \alpha_k\mu(-\tfrac{1}{\mu}g^k) \\ &= \sum_{s=0}^{k} \theta_s(-\tfrac{1}{\mu}g^s) \end{aligned}$$

for some convex combination $\{\theta_s\}_{s=0}^k$. (So $\theta_0, \ldots, \theta_k \geq 0$ and $\theta_0 + \cdots + \theta_k = 1$.) Using Jensen's inequality on this convex combination, we get

$$\begin{aligned} \mathbb{E}[\|g^k + \mu x^k\|^2] &\leq 2\mathbb{E}[\|g^k\|^2] + 2\mu^2\mathbb{E}[\|x^k\|^2] \\ &\leq 2(\|G^k\|^2 + \sigma^2) + 2\mu^2 \sum_{s=0}^{k} \theta_s \mathbb{E}\left[\left\| -\tfrac{1}{\mu}g^s \right\|^2\right] \\ &\leq 2(G^2 + \sigma^2) + 2(G^2 + \sigma^2) = 4(G^2 + \sigma^2) \end{aligned}$$

for $k = 0, 1, \ldots$.

Next,

$$
\begin{aligned}
&\mathbb{E}_k\left[\|x^{k+1} - x^\star\|_2^2\right] \\
&= \|x^k - x^\star\|_2^2 - 2\alpha_k\langle \mathbb{E}_k[g^k] + \mu x^k, x^k - x^\star\rangle + \alpha_k^2 \mathbb{E}_k[\|g^k + \mu x^k\|^2] \\
&\leq \|x^k - x^\star\|_2^2 - 2\alpha_k(F(x^k) - F(x^\star) + \frac{\mu}{2}\|x^k - x^\star\|^2) + \alpha_k^2 4(G^2 + \sigma^2) \\
&\leq (1 - \alpha_k\mu)\|x^k - x^\star\|_2^2 - 2\alpha_k(F(x^k) - F(x^\star)) + \alpha_k^2 4(G^2 + \sigma^2)
\end{aligned}
$$

for $k = 0, 1, \ldots$. Rearranging the terms and plugging in $\alpha_k = \frac{1}{\mu(k+1)}$,

$$
F(x^k) - F(x^\star) \leq \frac{\mu k}{2}\|x^k - x^\star\|_2^2 - \frac{\mu(k+1)}{2}\mathbb{E}_k\left[\|x^{k+1} - x^\star\|_2^2\right] + \frac{2(G^2 + \sigma^2)}{\mu(k+1)}
$$

With a telescoping sum argument, we have

$$
\sum_{s=0}^{k} F(x^s) - F(x^\star) \leq \frac{2(G^2 + \sigma^2)}{\mu}\sum_{s=0}^{k}\frac{1}{s+1} \leq \frac{2(G^2 + \sigma^2)}{\mu}(1 + \log(k+1)).
$$

Finally, we conclude with Jensen's inequality. □