First-order algorithms for black-box convex optimization

The Condition Number

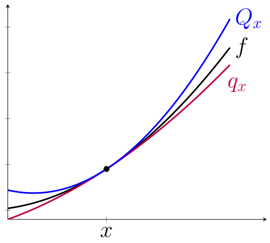
$$\min_{x \in \mathbf{E}} \; f(x),$$

where f is a β-smooth and α-strongly convex with $\alpha, \beta \geq 0$.
This implies f satisfies

$$\frac{\alpha}{2}\|y-x\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y-x \rangle \leq \frac{\beta}{2}\|y-x\|^2 \;\; \forall\, x, y \in \mathbf{E}.$$

Geometrically, f is sandwiched between the two quadratics

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



The Condition Number

The ratio

$$\kappa := \beta/\alpha$$

is called the *condition number* of f.

Intuitively, κ measures the "scaling" of the problem.
In particular, equality $\kappa = 1$ holds if and only if f is a spherical quadratic.

The condition number κ strongly influences convergence guarantees of numerical methods.

Throughout, we assume that f has at least one minimizer, which is automatic if $\alpha > 0$.
The symbols $\bar{f}$ and $\bar{x}$ will denote the minimal value of f and an arbitrary minimizer of f, respectively.

Gradient Descent

Lemma (Descent)

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

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

Proof.

Using

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

we have

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

as claimed.

Choice of η

Choosing η to be the maximizer of the concave quadratic $\eta \mapsto \eta\left(1 - \frac{\eta\beta}{2}\right)$ yields $\eta = \frac{1}{\beta}$.

This value of η in the Lemma shows that Gradient Descent Algorithm satisfies

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2\beta}\|\nabla f(x_t)\|^2 \qquad \text{for all } t \geq 0.$$

The Gradient Descent Algorithm (GDA):
Input: Starting point $x_0 \in \mathbf{E}$, parameter $\beta > 0$, iteration $T \in \mathbb{N}$.
Step $t = 0, 1, \ldots, T-1$:
Set $x_{t+1} = x_t - \frac{1}{\beta}\nabla f(x_t)$

Convergence of The Gradient Descent Algorithm

Theorem: Let $f\colon \mathbf{E} \to \mathbf{R}$ be β-smooth cvx. Then the iterates generated by GDA satisfy

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

Proof:

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

Convergence of The Gradient Descent Algorithm

We have

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

Subtract $\bar{f}$ from both sides and sum for $i = 0, \ldots, t-1$, the terms on the right side telescope, yielding

$$\sum_{i=0}^{t-1}(f(x_{i+1}) - \bar{f}) \leq \frac{\beta}{2}\sum_{i=0}^{t-1}\left(\|x_i - \bar{x}\|^2 - \|x_{i+1} - \bar{x}\|^2\right) \leq \frac{\beta}{2}\|x_0 - \bar{x}\|^2.$$

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

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

as claimed.

Convergence of GCA under Strong Convexity

Theorem: Let $f\colon \mathbf{E} \to \mathbf{R}$ be α-strongly cvx and β-smooth. Then the iterates generated by GDA satisfy

$$(\star) \qquad f(x_{t+1}) - \bar{f} \leq \left(1 - \frac{1}{2\kappa}\right)(f(x_t) - \bar{f}),$$

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

Proof: To see $(\star)$, we combine the inequality

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

with

$$(\diamond) \qquad f(x) - \bar{f} \leq \tfrac{1}{\alpha}\|v\|^2 \qquad \text{for all } x \in \mathbf{E},\ v \in \partial f(x)$$

to deduce

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

Adding and subtracting $\bar{f}$ from the left-side yields

$$(f(x_{t+1}) - \bar{f}) - (f(x_t) - \bar{f}) \leq -\frac{1}{2\kappa}(f(x_t) - \bar{f}).$$

Rearranging completes the proof of $(\star)$.

Convergence of GCA under Strong Convexity

Next, observe

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

where (**??**) follows from strong convexity. ($\diamond$) shows the second term in (**??**) is nonpositive, and therefore the quantity $\|x_{t+1} - \bar{x}\|^2$ tends to zero at the linear rate $1 - \kappa^{-1}$.

Convergence of GCA under Strong Convexity

To obtain $(\star\star)$ note that strong convexity and $(\diamond)$ guarantees that

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

and therefore

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

Combining this estimate with (**??**) and rearranging yields $(\star\star)$.

Accelerated gradient descent

Accelerated Gradient Descent Algorithm (AGDA)
Input: Starting point $x_0 \in \mathbf{E}$. Set $t = 0$ and $a_0 = a_{-1} = 1$, $x_{-1} = x_0$

For: $t = 0, \ldots, T$

Set

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

AGDA convergence rate

Theorem: Let $f\colon \mathbf{E} \to \mathbf{R}$ be a convex and β-smooth function. Then the iterates $\{x_t\}$ generated by AGDA satisfy

$$f(x_t) - \bar{f} \leq \frac{2\beta\|\bar{x} - y_0\|^2}{(t+1)^2}.$$

Subgradient method for nonsmooth convex minimization

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

where $f : \mathbf{E} \to \mathbf{R}$ is cvx and L-Lipschitz on an neighborhood of a closed convex set $Q \subset \mathbf{E}$.

Subgradient method for nonsmooth convex minimization

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

where $f \colon \mathbf{E} \to \mathbf{R}$ is cvx and L-Lipschitz on an neighborhood of a closed convex set $Q \subset \mathbf{E}$.

Algorithm: The Projected Subgradient Algorithm (PSGA)
Input: Initial $x_0 \in \mathbf{E}$, iteration $T \in \mathbb{N}$, sequence $\{\eta_t\} \subset (0, \infty)$.
Step $t = 0, 1, \ldots, T-1$:
  Choose $v_t \in \partial f(x_t)$
  Set $x_{t+1} = \mathrm{proj}_Q(x_t - \eta_t v_t)$

PSGA Convergence

Let $f\colon \mathbf{E} \to \mathbf{R}$ be cvx and L-Lipschitz continuous on a neighborhood of a closed convex set $Q \subset \mathbf{E}$. Then the iterates generated by PSGA satisfy

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

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

$$f\left(\frac{1}{T+1} \sum_{t=0}^{T} x_t\right) - \bar{f} \le \frac{RL}{\sqrt{T+1}}.$$

PSGA Convergence

Proof:

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

Iterating the recursion yields

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

implying

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

Convexity gives

$$f\left(\tfrac{1}{\sum_{t=0}^T \eta_t} \textstyle\sum_{t=0}^T \eta_t x_t\right) - \bar{f} \le \frac{\sum_{t=0}^T \eta_t (f(x_t) - \bar{f})}{\sum_{i=0}^t \eta_t}.$$

and combining gives

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

PSGA Convergence under strong convexity

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

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

PSGA Convergence under strong convexity

Proof: We have already shown that

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

Applying L=Lip. cont. and α-strong cvxity gives

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

Rearrange and divide by $2\eta_t$ to get

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

Plug in $\eta_t := \frac{2}{\alpha(t+1)}$ and multiply by t,

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

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

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

Divide by $\sum_{i=1}^t i = \frac{t(t+1)}{2}$ and use cvxity to conclude

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