Lecture 1 Gradient Descent

- idea behind gradient descent
- smoothness
- strong convexity
- convergence analysis of GD under smoothness and strong convexity / smothness assumptions

convex optimization problem

$$\min_{x \in C} f(x)$$

convex function

convex sets

let's assume for simplicity that $C = \mathbb{R}^n$

we start looking at a greedy approach

① start at x_0

② for $i = 0$ to ...

- find y s.t. $f(y) < f(x_i)$
- $x_i = y$

Question 1

Does it work?

$\Rightarrow$

Answer

As long as we can find y s.t $f(y) < f(x_i)$ yes it works. Because all local minimum are also global minima. (f is convex)

Question 2

How do we find y s.t $f(y) < f(x_i)$

(let's assume that everything is differentiable and fine)

since $f(y) \simeq f(x) + \nabla f(x)(y-x)$ if y is close to x

take $y-x = \varepsilon \cdot (-\nabla f(x))$ $\Longrightarrow$

$$\Rightarrow f(y) \simeq f(x) - \varepsilon \|\nabla f(x)\|_2^2 < f(x)$$

$$y = x + \varepsilon(-\nabla f(x))$$

step size

descending direction

(some algorithms use different directions v as long as $\langle v, \nabla f(x) \rangle < 0$)

Gradient Descent method (GD)

- start from x_0, $t=0$
- determine a descent direction v_t (e.g. $-\nabla f(x_t)$)
- choose a step size η_t can depend on t
- update $x_{t+1} = x_t + \eta_t \cdot v_t$

  $t \leftarrow t+1$

NO ⟹ repeat

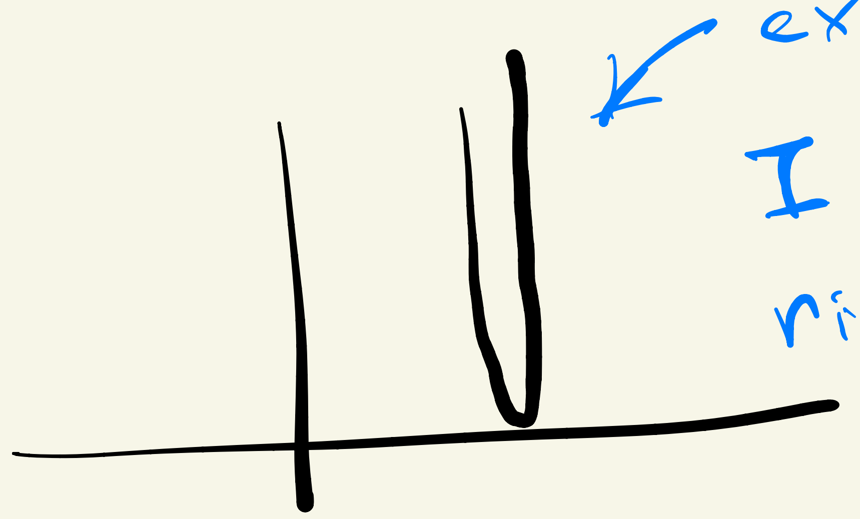
Is the stopping criterion satisfied? $\xrightarrow{\text{YES}}$ return x_{t+1}

Line search = choose the step size that minimizes the objective function along some fixed descent direction

why? Because it may be the case that $\min\limits_{\eta \in \mathbb{R}} f(x_t + \eta v_t)$ is easy!!!

Can we say something about the convergence rate without any additional assumptions?

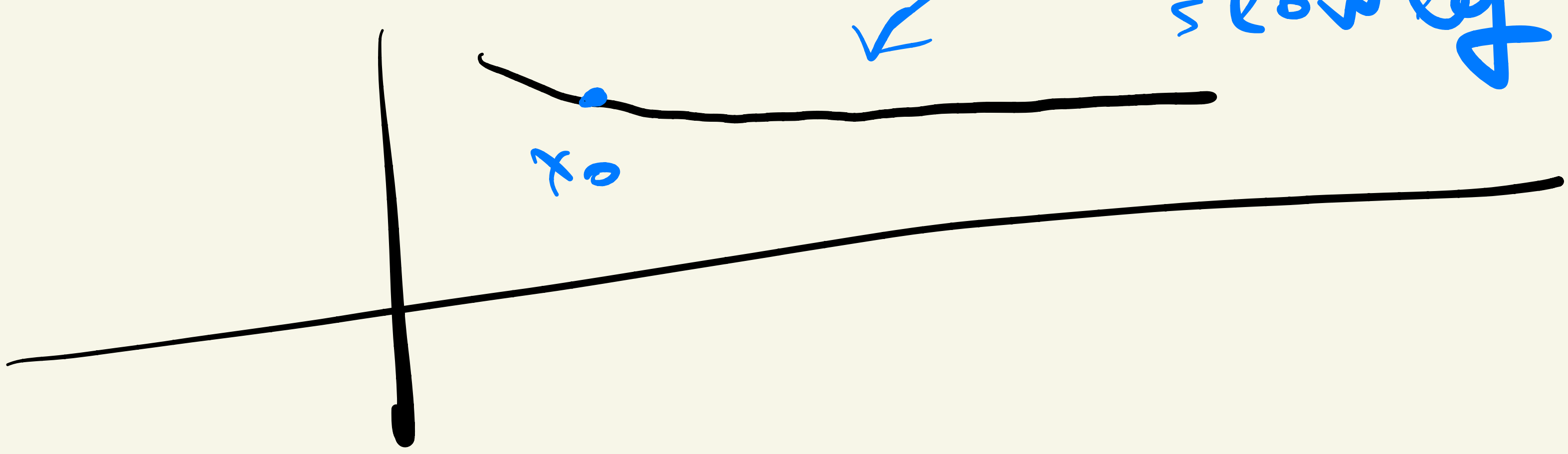
No



extremely steep.

I bounce left and right

decreases too slowly



Smoothness

for twice differentiable functions this is equivalent to $\nabla^2 f(x) \preceq \beta \cdot I$

Def

We say that a continuously differentiable function f is β-smooth if

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x-y\|$$

↳ the gradient does not change too quickly

Question.

When is the first order Taylor approximation $f(x) + \nabla f(x)(y-x)$ of $f(y)$ a good approximation?

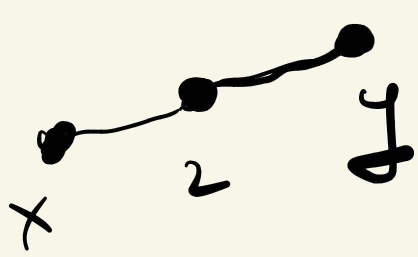
without any assumption

only if y is very close to x

when f is β-smooth and β is small

y can be more distant from x, because we know that the gradient changes slowly

elaborate more



is $f(x) + \nabla f(x)(y-x)$ a good approximation for y

$x - z$ are close

$$f(z) \simeq f(x) + \nabla f(x)(z - x)$$

$z - y$ are close

$$f(y) \simeq f(z) + \nabla f(z)(y-z)$$

$\oplus$

smoothness

$$\nabla f(x) \simeq \nabla f(z)$$

$$f(y) \simeq f(x) + \nabla f(x)(y-x)$$

Intuitive take away point

If $\|\nabla f(x)\|$ is large we can be aggressive in our optimization step

(we are searching for x^* s.t $\|\nabla f(x^*)\| = 0$ and $\nabla f(x)$ changes slowly)

Unconstrained case

Some claims about β-smooth functions

① $f(y) \leq f(x) + \nabla f(x)(y-x) + \frac{\beta}{2}\|y-x\|_2^2$

good linear approximation

proof

$\exists$ z in the line connecting x and y s.t

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

$$\leq f(x) + \nabla f(x)(y-x) + \frac{1}{2}(y-x)^T \cdot \beta \cdot I (y-x)$$

$$= f(x) + \nabla f(x)(y-x) + \frac{\beta}{2}\|y-x\|_2^2$$

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

small β
① I can be very aggressive
② I decrease a lot

taking the descent direction of the negative gradient and a step size of $1/\beta$

I am sure I am gonna decrease !!!

proof

① $\Rightarrow f(y) \leq f(x) + \nabla f(x)(y-x) + \frac{\beta}{2}\|y-x\|_2^2 \Rightarrow$

$$\Rightarrow f(y) - f(x) \leq \nabla f(x)(y-x) + \frac{\beta}{2}\|y-x\|^2$$

minimize the UB

$$\nabla_y(UB) = \nabla f(x) + \beta(y-x) \Rightarrow$$

$$\Rightarrow \nabla_y(UB) = 0 \Rightarrow y = x + \left(\frac{1}{\beta}\right)\left(-\nabla f(x)\right)$$

by replacing we get

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

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

③ $f(y) \geq f(x) + \nabla f(x)(y-x) + \frac{1}{2B}\|\nabla f(x)-\nabla f(y)\|_2^2$

a "little bit" better convexity.

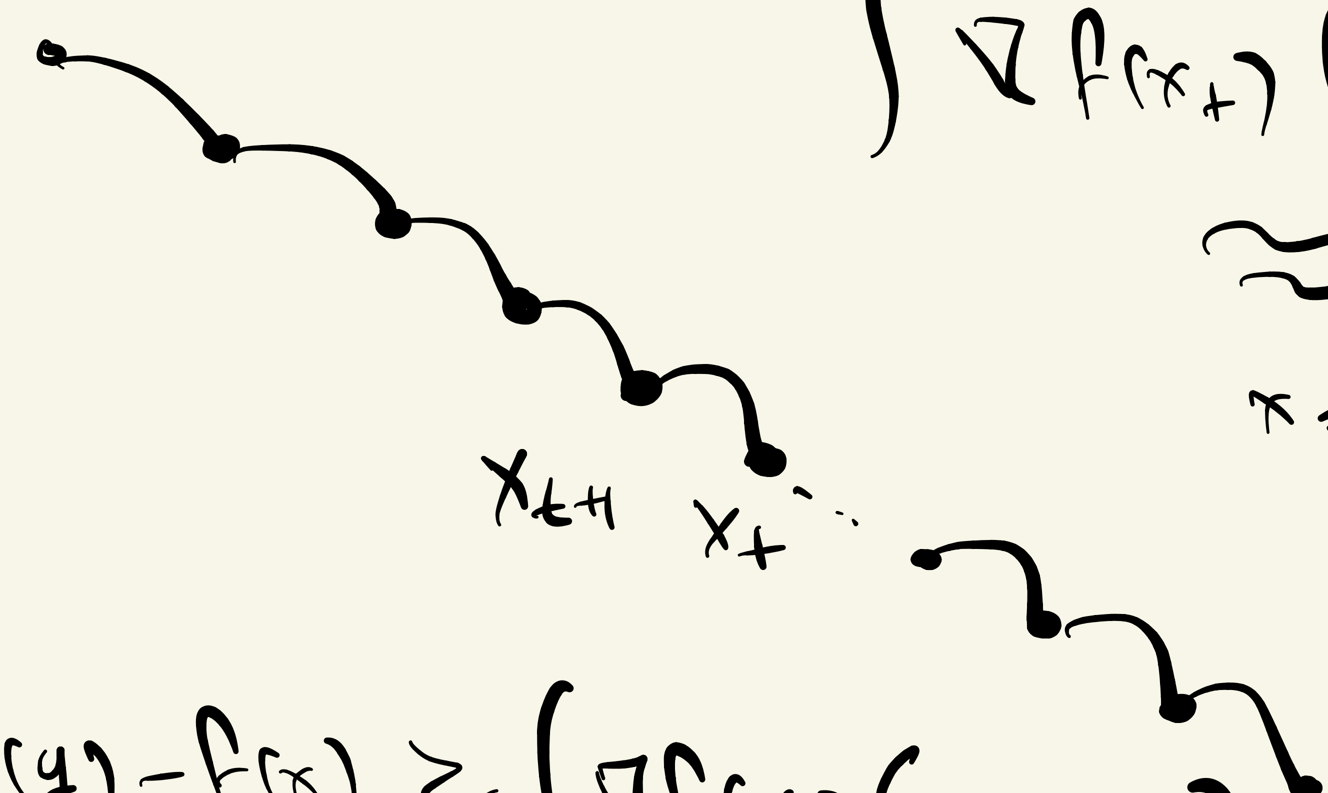
Intuition: I increase by consecutive linear approximation) by claiming that $\nabla f(x)$ is fixed

$f(y)$

$$\int \nabla f(x_t)(x_{t+1}-x_t)$$

$$\approx \nabla f(x_0)(x_{t+1}-x_0)$$

x_0 fixed



$$f(y)-f(x) \geq \int \nabla f(x_t)(x_{t+1}-x_t)$$

$$\approx \nabla f(y)\left(y-\frac{x+y}{2}\right) + \nabla f(x)\left(\frac{x+y}{2}-x\right) =$$

$$= \nabla f(y)\left(\frac{y-x}{2}\right) - \nabla f(x)\left(\frac{y-x}{2}\right) =$$

$$= (\nabla f(y)-\nabla f(x))\left(\frac{y-x}{2}\right) \approx$$

lets assume that because of convexity

$$\|\nabla f(y)-\nabla f(x)\|\|y-x\| \geq$$

$$\geq \|\nabla f(y)-\nabla f(x)\|^2/2B$$

proof

for any z we have

$$f(x) + \nabla f(x)(z-x) \leq f(z) \leq f(y) + \nabla f(y)(z-y) + \frac{b}{2}\|z-y\|^2$$

$y-x+z-y$

$$\Rightarrow f(y) \geq f(x) + \nabla f(x)(\widetilde{z-x}) - \nabla f(y)(z-y) - \frac{b}{2}\|z-y\|^2$$

$-(\nabla f(y) - \nabla f(x))(z-y)$

(*)

$$f(y) \geq f(x) + \nabla f(x)(y-x) + \overbrace{\nabla f(x)(z-y) - \nabla f(y)(z-y)} - \frac{b}{2}\|z-y\|^2 =$$

maximize $(\nabla f(y) - \nabla f(x))(z-y) + \frac{b}{2}\|z-y\|^2$

$$\nabla_z(\ ) = 0 \Rightarrow (\nabla f(y) - \nabla f(x)) + bz - by = 0$$

$$\Rightarrow z = y - \frac{1}{b}(\nabla f(y) - \nabla f(x))$$

replace $z-y = -\frac{1}{b}(\nabla f(y) - \nabla f(x))$ to

(*)

corolarry (coercivity of the gradient)

$$(\nabla f(y) - \nabla f(x))(y-x) \geq \frac{1}{\beta} \|\nabla f(y) - \nabla f(x)\|_2^2$$

proof

from a "little bit" better convexity.

$$f(y) - f(x) \leq \nabla f(x)(x-y) - \frac{1}{2\beta} \|\nabla f(y) - \nabla f(x)\|^2$$

$$f(x) - f(y) \leq \nabla f(y)(y-x) - \frac{1}{2\beta} \|\nabla f(y) - \nabla f(x)\|^2$$

$\oplus \Rightarrow \ldots$

Strong convexity

$\nabla^2 f(x) \succeq \alpha \cdot I$

f twice differentiable

Def

we say that f is α-strongly convex

if $f(y) \geq f(x) + \nabla f(x)(y-x) + \frac{\alpha}{2}\|y-x\|^2$

in β-smoothness we have $\frac{1}{2\beta}\|\nabla f(y) - \nabla f(x)\|_2^2$

Intuitive take-away point

when $\|\nabla f(x)\|$ is small

stop. (because $\|y-x\|^2$ dominates the RHS

and consequently $f(y) \geq f(x)$ if we move

slightly move) → optimal solution is closer

claims

① $\frac{1}{2\alpha}\|\nabla f(x)\|_2^2 \geq f(x) - f(x^*)$

If gradient is small we are good!!!

proof

$$f(y) \geq f(x) + \nabla f(x)(y-x) + \frac{\alpha}{2}\|y-x\|^2$$

$$\Rightarrow f(x) - f(y) \leq \underbrace{-\frac{\alpha}{2}\|y-x\|^2 - \nabla f(x)(y-x)}_{\text{maximize}}$$

$$\nabla_y (UB) = 0 \Rightarrow -\alpha(y^* - x) - \nabla f(x) = 0$$

$$\Rightarrow y^* = x - \frac{1}{\alpha}\nabla f(x)$$

$$\Rightarrow f(x) - f(y) \leq \frac{1}{2\alpha}\|\nabla f(x)\|^2$$

holds $\forall y \Rightarrow y = x^*$ ← optimal solution

② $\|x - x^*\|_2 \leq \frac{2}{\alpha} \|\nabla f(x)\|_2^2$

If the gradient is small not only $f(x) \simeq f(x^*)$ but x is actually close to x^*

proof

$f(x^*) \geq f(x) + \nabla f(x)(x^* - x) + \frac{\alpha}{2}\|x - x^*\|_2^2$

$\geq f(x) - \|\nabla f(x)\| \|x^* - x\| + \frac{\alpha}{2}\|x^* - x\|$

cauchy-schwartz

$f(x^*) - f(x) \leq 0$ (x^* is a minimizer

=> ②

---

other common assumptions

L-Lipschitz $|f(x) - f(y)| \leq L \cdot \|x - y\|$

my function does not change very fast

Diameter $\forall x, y \in C \quad \|x - y\| \leq R$

I am always at most R far away from x^*

convergence analysis of GD

1. a-strongly convex $\quad (b > a > 0)$
   b-smooth

$$x_{t+1} = x_t - \frac{1}{b} \nabla f(x)$$

then $f(x_T) - f(x^*) \le \left(1 - \frac{a}{b}\right)^T \left(f(x_0) - f(x^*)\right)$

$e^x \geq 1+x, \forall x$

$$\le e^{-\frac{aT}{b}} \left(f(x_0) - f(x^*)\right) \le$$

To ensure

$$\le e^{-\frac{aT}{b}} \left(\frac{b}{2}\right) \|x_0 - x^*\|_2^2$$

$f(x_T) - f(x^*) < \varepsilon$

b-smoothness definition and $\nabla f(x^*) = 0$

$$T = \frac{b}{a} \ln\left(\frac{f(x_0) - f(x^*)}{\varepsilon}\right)$$

$$f(x_0) \le f(x^*) + \nabla \cancel{f(x^*)}^{=0} (x^* - x) + \frac{b}{2} \|x^* - x\|_2^2$$

condition number (happy when it is small)

$$O(\log 1/\varepsilon)$$

Proof $f(x_T) - f(x^*) \le \left(1 - \frac{\alpha}{\beta}\right)^T \left(f(x_0) - f(x^*)\right)$

Idea

In every iteration $x_t \to x_{t+1}$

① we decrease our objective by

$$\frac{1}{2\beta} \|\nabla f(x_t)\|^2$$

② we are at most $\frac{1}{2\alpha} \|\nabla f(x_{t+1})\|^2$ far away from the optimal solution

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

$$\Rightarrow f(x_{t+1}) - f(x^*) \le f(x_t) - f(x^*) - \frac{1}{2\beta} \underbrace{\|\nabla f(x_t)\|^2}_{\ge 2\alpha (f(x_t) - f(x^*))}$$

$$\Rightarrow f(x_{t+1}) - f(x^*) \le \left(1 - \frac{\alpha}{\beta}\right)\left(f(x_t) - f(x^*)\right)$$

...

② only β-smoothness

$$x_{t+1} = x_t - \frac{1}{\beta} \nabla f(x_t)$$

then $f(x_T) - f(x^*) \leq \frac{16\beta \|x_0 - x^*\|^2}{T}$

proof

(helpful notation $\delta_t = f(x_t) - f(x^*)$)

$$\delta_{t+1} - \delta_t \leq -\frac{1}{2\beta}\|\nabla f(x)\|^2$$

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2\beta}\|\nabla f(x_t)\|^2$$ (by the smoothness assumption) claim ② for smooth functions

by first order condition

$$f(x_t) - f(x^*) \leq \nabla f(x_t)(x_t - x^*) \leq$$

$$\leq \|\nabla f(x_t)\| \, \|x_t - x^*\| \Rightarrow$$

$$\|\nabla f(x_t)\| \geq \frac{f(x_t) - f(x^*)}{\|x_t - x^*\|}$$

decreases

let's be optimistic and assume that decreases

then $\|\nabla f(x_t)\| \geq \frac{f(x_t) - f(x^*)}{\|x_0 - x^*\|} = c \cdot \delta_t$

$$\delta_{t+1} - \delta_t \leq -\frac{c^2 \delta_t^2}{2\beta}$$

$\left(c = \frac{1}{\|x_0 - x^*\|}\right)$

$$\delta_{t+1} - \delta_t \leq -\frac{c^2}{2B} \delta_t^2$$

how much time I do need to half the distance $\delta_{t'} \simeq \frac{\delta_t}{2}$?

Before I get $\delta_t' < \frac{\delta_t}{2}$

I have a decrease of at least

$$\frac{c^2}{2B} \frac{\delta_t^2}{4} = \frac{c^2 \delta_t^2}{8B} \text{ at each step}$$

$$\Rightarrow T \cdot \frac{c^2 \delta_t^{\cancel{2}}}{8B} < \frac{\cancel{\delta_t}}{2} \Rightarrow T < \frac{4B}{c^2 \delta_t}$$

If I need to half my error J times $\Rightarrow$

$$\Rightarrow T_{total} < \frac{4B}{c^2 \delta_0} + \frac{4B}{c^2} \cdot \left(\frac{2}{\delta_0}\right) + \dots + \frac{4B}{c^2} \cdot \frac{2^J}{\delta_0} =$$

$$= O\left(\frac{4B}{c^2} \cdot \frac{2^{J+1}}{\delta_0}\right) \Rightarrow c = \frac{1}{\|x^* - x_0\|^2}$$

$$\Rightarrow \frac{\delta_0}{2^J} = O\left(\frac{B \cdot \|x^* - x_0\|^2}{T_{total}}\right) \Rightarrow f(x_{T_{total}}) - f(x^*) = O\left(\frac{B \cdot \|x^* - x_0\|^2}{T_{total}}\right)$$

Last claim to prove

$\|x_t - x^*\|^2$ is non decreasing

Proof

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

$$= \|x_t - x^*\| + \frac{1}{\beta^2}\|\nabla f(x_t)\|^2$$

$$- \frac{2}{\beta}\nabla f(x_t)(x_t - x^*) \leq$$

$$\leq \|x_t - x^*\|^2 - \frac{1}{\beta^2}\|\nabla f(x_t)\|^2$$

by coercivity and $\nabla f(x^*) = 0$

$$\left((\nabla f(y) - \nabla f(x))(y-x) \geq \frac{1}{\beta}\|\nabla f(y) - \nabla f(x)\|_2^2 \right)$$

Now we summarize the performance of GD for different properties of the convex f that we are optimizing over, and we compare it to Nesterov accelerated method together with lower bounds on the performance of any optimization algorithm.

# iterations to get an ϵ-additive error	GD	Nesterov's accelerated method	λB
L-Lipschitz convex f	$O\left(\frac{L^2}{\epsilon^2}\right)$		$O\left(\frac{L^2}{\epsilon^2}\right)$
b-smooth convex f	$O\left(\frac{b}{\epsilon}\right)$	$O\left(\sqrt{\frac{b}{\epsilon}}\right)$	$O\left(\sqrt{\frac{b}{\epsilon}}\right)$
b-smooth and a-strongly convex f	$O\left(\frac{b}{a}\log\frac{1}{\epsilon}\right)$	$O\left(\sqrt{\frac{b}{a}}\log\frac{1}{\epsilon}\right)$	$O\left(\sqrt{\frac{b}{a}}\log\frac{1}{\epsilon}\right)$

we hide the dependence on the initial distance from the optimum ($\|x_0 - x^*\|$) in the $O(\cdot)$ notation. Here we focus mainly on the dependence with ϵ.

Although, as we will see for the min-cut problem, getting a better dependence w.r.t. $\|x^* - x_0\|$ is important, and Nesterov's accelerated method also achieves that.

For a table summarizing many more convergence rates of various optimization algorithms. Please see the end of chapter 1 in Bubeck's book (Convex Optimization: Algorithms and Complexity)