Lecture 13 Interior point method

- steepest descent
- Newton method, convergence analysis when "close" to the optimum
- logarithmic Barrier functions
- central path
- Barrier method

---

steepest descent

$$f(x+v) \approx f(x) + \langle \nabla f(x), v \rangle$$

- Gradient descent selects $v = -\eta \nabla f(x)$ as a descent direction
- Mirror descent tries to minimize the first-order approximation + a penalty term

not to go far away in terms of a certain distance function

$$x_{t+1} = \underset{x \in C}{\operatorname{argmin}} \left\{ f(x_t) + \langle \nabla f(x_t), \underbrace{x - x_t}_{= v} \rangle + \frac{1}{\eta_t} D_\varphi(x, x_t) \right\}$$

- Steepest descent intuition is very similar to Mirror descent's intuition. We choose a descent direction v that minimizes the first-order approximation constraining v to belong in the unit ball of an arbitrary norm.

$$v_t = \arg\min_v \{ f(x_t) + \langle \nabla f(x_t), v \rangle \mid \|v\| \leq 1 \} =$$
$$= \arg\min_v \{ \langle \nabla f(x_t), v \rangle \mid \|v\| \leq 1 \}$$
$$x_{t+1} = x_t + v_t$$

Some examples

- If we use the ℓ_2-norm then $v_t = -\dfrac{\nabla f(x_t)}{\|\nabla f(x_t)\|_2}$ (as we would expect) the negative gradient direction
- If we use the ℓ_1-norm then $v_t = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ position i such that $|(\nabla f(x))_i|$ is maximized (coordinate descent)
- If we use a quadratic norm $\|x\|_Q = \sqrt{x^T Q x}$ then

$$v_t = \arg\min_v \{ \langle \nabla f(x_t), v \rangle \mid \|v\|_Q \leq 1 \} =$$
$$= \arg\min_v \{ \langle \nabla f(x_t), v \rangle \mid \|Q^{1/2} v\|_2 \leq 1 \} =$$
$$= \arg\min_v \{ \langle (Q^{-1/2})^T \nabla f(x), Q^{1/2} v \rangle \mid \|Q^{1/2} v\|_2 \leq 1 \}$$
$$= Q^{-1/2} \arg\min_y \{ \langle Q^{-1/2} \nabla f(x), y \rangle \mid \|y\|_2 \leq 1 \} =$$
$$= Q^{-1/2} (-Q^{1/2} \nabla f(x)) / \|Q^{-1/2} \nabla f(x)\|_2 =$$

$$= Q^{-1/2}\left(-Q^{1/2}\nabla f(x)\right)/\|Q^{-1/2}\nabla f(x_t)\|_2 =$$

$$= -Q^{-1}\nabla f(x_t)/\|Q^{-1/2}\nabla f(x_t)\|_2 =$$

$$= -Q^{-1}\nabla f(x_t)/\|\nabla f(x_t)\|_{Q^{-1}} =$$

$$= \frac{Q^{-1}\nabla f(x)}{\|\nabla f(x)\|_*}$$ (the dual norm of $\|\cdot\|_Q$ is $\|\cdot\|_{Q^{-1}}$)

Newton method

use the hessian to penalize how far we go in certain directions.

why?

(1) Because the hessian is a good guess of how the geometry around the point that we are now looks like

(2) not going to far in terms of $\sqrt{x^T H x}$ let's the first order approximation be valid

Newton step

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

full Newton step

If we normalize choosing a step-size of $\frac{1}{\|\nabla f(x_t)\|_{(\nabla^2 f(x_t))^{-1}}}$ then we get the steepest descent update using $\|\cdot\|_{\nabla^2 f(x_t)}$

The Newton step can be equivalently defined as the direction which minimizes the second order approximation.

$$f(x+v) \simeq f(x) + \langle \nabla f(x), v \rangle + \frac{1}{2} v^T \nabla^2 f(x) v$$

$$\nabla_u \left(f(x) + \langle \nabla f(x), v \rangle + \frac{1}{2} v^T \nabla^2 f(x) v \right) = 0 \Rightarrow$$

$$\Rightarrow \nabla f(x) + \nabla^2 f(x) \cdot v = 0 \Rightarrow v = -(\nabla^2 f(x))^{-1} \nabla f(x)$$

Convergence analysis

Without a carefull selection of stepsize the Newton method may even diverge. The important property about the Newton method that we are going to prove is that if the starting point x_0 is sufficiently close to the optimum x^* then the convergence is quadratic

ie $$\lim_{n \to +\infty} \frac{\|x_{t+1} - x^*\|_2}{\|x_t - x^*\|_2^2} = \Theta(1)$$

to that end let's introduce some additional notation

$$\Delta x_t = x_{t+1} - x_t = -\left(\nabla^2 f(x_t)\right)^{-1} \nabla f(x_t)$$

Newton decrement

$$\lambda(x_t) = \left((\nabla f(x_t))^T (\nabla^2 f(x_t))^{-1} \nabla f(x_t) \right)^{1/2}$$

$$\frac{1}{2}\lambda^2(x_t) = f(x_t) - \inf_v \left(f(x_t) + \langle \nabla f(x_t), v \rangle + \frac{1}{2} v^T \nabla^2 f(x_t) v \right)$$

Theorem (quadratic convergence close to the optimum)

f has a M-Lipschitz Hessian, i.e.

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq M\|x-y\|_2$$

↳ operator norm: $\max_x \frac{\|Ax\|_2}{\|x\|_2}$ = maximum eigenvalue of A

equivalent to

$A \preceq B \iff B - A \in S^n_+$

$$-\|x-y\|_2 M \cdot I \preceq \nabla^2 f(x) - \nabla^2 f(y) \preceq \|x-y\|_2 \cdot M \cdot I$$

In words → the second order approximation is locally good.

Let x^* be the optimum and $\nabla^2 f(x^*) \succeq \mu I$, $\mu > 0$.

→ the smaller μ, the flatter is the landscape close to the optimum

If $\|x_0 - x^*\|_2 \leq \mu/_{2M}$ then

$$\|x_{t+1} - x^*\|_2 \leq \frac{M}{\mu} \|x_t - x^*\|_2^2$$

bigger M = less information

⇒ we need to be closer

⇒ the closer we have to be to converge fast.

proof

$$x_{t+1} - x^* = x_t - x^* - \left(\nabla^2 f(x_t)\right)^{-1} \nabla f(x_t)$$

to relate $\nabla f(x_t)$ with the Hessian $\nabla^2 f(x_t)$ we use the formular

$$\nabla f(x+v) - \nabla f(x) = \int_0^1 \nabla^2 f(x+sv) v \, ds$$

$x = x^*$
$\Longrightarrow$
$x + v = x_t$

$$\nabla f(x_t) = \int_0^1 \nabla^2 f(x^* + s(x_t - x^*)) v \cdot ds$$

$\Rightarrow v = x_t - x^*$

$\nabla f(x^*) = 0$

Thus $x_{t+1} - x^* =$

$$= x_t - x^* - \left(\nabla^2 f(x_t)\right)^{-1} \int_0^1 \nabla^2 f\left(x^* + s(x_t - x^*)\right)(x_t - x^*) \, ds$$

$$= \left(\nabla^2 f(x_t)\right)^{-1} \int_0^1 \left(\nabla^2 f(x_t) - \nabla^2 f(x^* + s(x_t - x^*))\right)(x_t - x^*) \, ds$$

$\Rightarrow \|x_{t+1} - x^*\|_2 \le$

$$\le \left\|\left(\nabla^2 f(x_t)\right)^{-1}\right\|_2 \left(\int_0^1 \| \quad \|_2 \, ds\right) \|x_t - x^*\|_2 \Rightarrow$$

$\Rightarrow$

$$\|x_{t+1}-x^*\|_2 \le \|(\nabla^2 f(x_t))^{-1}\|_2 \left(\int_0^1 \mu \cdot \|(1-s)(x_t-x^*)\|_2 ds\right) \cdot \|x_t-x^*\|_2$$

$$\le \|\nabla^2 f(x_t)^{-1}\|_2 \cdot \frac{\mu}{2} \|x_t - x^*\|_2^2$$

It only remains to upper bound it by $\frac{1}{\mu}$.

using induction

$$\|x_t-x^*\|_2 \le$$
$$\le \|x_0-x^*\|_2 \le$$
$$\le \frac{\mu}{2M}$$

$$(\nabla^2 f(x_t)) \succeq \mu \cdot I$$

$$\nabla^2 f(x_t) \succeq \nabla^2 f(x^*) - M\|x_t-x^*\|_2 I \succeq$$

$$\succeq \mu \cdot I - M \frac{\mu}{2M} I = \frac{\mu}{2} I$$

Logarithmic barrier functions

$$\min f_0(x)$$
$$\text{s.t } f_i(x) \le 0 \quad \forall i \in [m]$$

$$\xrightarrow{\text{unconstrained}} \min \; t f(x) - \sum_{i=1}^{m} \log(-f_i(x))$$

penalty function

t: relative importance of the penalty function.

Pros

- $t \to +\infty$ then the two problems are equivalent

- $-\log(-f_i(x))$ is convex if f_i is convex

- $\varphi(x) = -\sum_{i=1}^{m} \log(-f_i(x))$ is a convex differentiable function.

$$\nabla \varphi(x) = \sum_{i=1}^{m} -\frac{\nabla f_i(x)}{f_i(x)}$$ and

$$\nabla^2 \varphi(x) = \sum_{i=1}^{m} \left(\frac{\nabla f_i(x) (\nabla f_i(x))^T}{f_i^2(x)} - \frac{\nabla^2 f_i(x)}{f_i(x)} \right)$$

Cons

the minimizer of the two problems may not be the same

Central path

Consider the problem

$$\min \; t f_0(x) + \varphi(x)$$
$$\text{s.t} \quad Ax = b$$

Let $x^*(t)$ be the unique optimal solution for $t>0$. The path $\{x^*(t)\}_{t>0}$ is called the central path.

optimality conditions:

- $A x^*(t) = b$
- $f_i(x^*(t)) < 0 \;\; \forall\, i \in [m]$
- $t \nabla f_0(x^*(t)) + \nabla \varphi(x^*(t)) + \beta^T A = 0$

$$\Rightarrow \nabla f_0(x^*(t)) - \sum_{i=1}^{m} \frac{\nabla f_i(x^*(t))}{t f_i(x^*(t))} + \frac{\beta^T A}{t} = 0$$

for some β

$$\Rightarrow \lambda_i = -\frac{1}{t f_i(x^*(t))}, \quad \mu_i = \frac{\beta_i}{t} \quad \forall\, i \in [m]$$

$$\nabla f_0(x^*(t)) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x^*(t)) + \mu^T A = 0 \Longrightarrow$$

Thus $x^*(t)$ is the minimizer of

$$g(\lambda,\mu) = \inf_x f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \mu^T(Ax-b)$$

which is a lower bound (dual problem)

of the initial problem $\min f_0(x)$

$\quad$ s.t $f_i(x) \leq 0$

$\quad \forall i \in [m]$

$\quad Ax = b$

$$f_0(x^*(t)) - g(\lambda,\mu) = -\sum_{i=1}^m \lambda_i f_i(x^*(t)) + \mu^T(Ax^*(t)-b)$$

$$= \frac{m}{t} \quad \left(\lambda_i = \frac{-1}{t f_i(x^*(t))} \quad , Ax^*(t) = b \right)$$

Conclusion

Let p^* be the optimal value of the original problem. Then

$$f_0(x^*(t)) - p^* \leq m/t$$

$\Rightarrow t = m/\varepsilon$ and by solving the unconstrained problem we get an ε-additive error.

(problem $\Rightarrow$ it is too slow)

Barrier method

1) start with a small t,
2) compute $x^*(t)$
3) increase t and use the previous solution as the starting point of the new unconstrained problem

why? If $x^*(t)$ is close to $x^*(t_{new})$ then Newton method converges really fast and step (3) is consequently super fast

of course t_{new} has to be

- small enough to guarantee quadratic convergence
- big enough to arrive at $t > m/\varepsilon$ in a few iterations

Barrier method algorithm

find a strictly feasible solution x

set $t = t^{(0)} > 0$

repeat until $t \geq m/\varepsilon$

(1) compute $x^*(t)$ by minimizing

$$tf_0 + \phi \quad \text{using } x_0 = x$$
$$\text{s.t. } Ax = b$$

(2) $x = x^*(t)$

(3) $t = a \cdot t \quad , a > 1$

It will turn out (next lecture) that $a = 1 + 1/\sqrt{m}$ works and hence at most $\sqrt{m} \log \frac{m}{\varepsilon}$ iterations suffice.