Two Important Isolated Results

1. Stochastic Gradient Descent

Often, especially in machine learning, you get a **stochastic** estimate of the gradient.

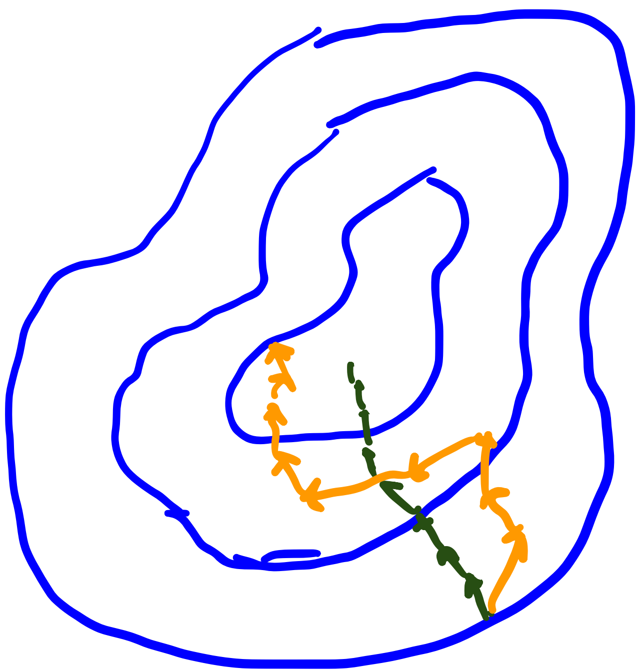
$$\underset{\theta}{\text{minimize}} \sum_{(x,y)\in\mathcal{D}} \text{loss}(m(x;\theta), y)$$

$$f(\theta) = \sum_{(x,y)\in\mathcal{D}} \text{loss}(m(x;\theta), y)$$

$$\nabla_\theta f = \sum_{(x,y)\in\mathcal{D}} \nabla_\theta \, \text{loss}(m(x;\theta), y)$$

$$\widehat{\nabla_\theta f} = \sum_{(x,y)\in\mathcal{B}} \nabla_\theta \, \text{loss}(m(x;\theta), y)$$

$$\mathcal{B} \subset \mathcal{D}$$



Stochastic Gradient Descent

loop:

$$\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \alpha_k \widehat{\nabla_{\mathbf{x}} f}$$

1. Stochastic Gradient Descent

Stochastic Gradient Descent

loop:

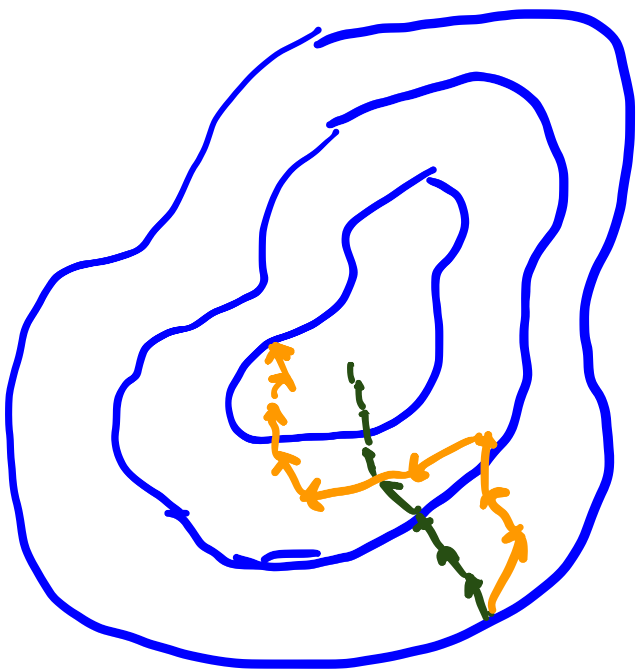
$$\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \alpha_k \widehat{\nabla_{\mathbf{x}} f}$$

Convergence Guarantee (roughly):

If these assumptions are satisfied:

1. $E\left[\widehat{\nabla f}\right] = \nabla f$
2. $\sum_{k=1}^{\infty} \alpha_k = \infty$ (steps not too small)
3. $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$ (steps not too big)
4. Other very mild assumptions

e.g. $\alpha_k = \frac{1}{k}$



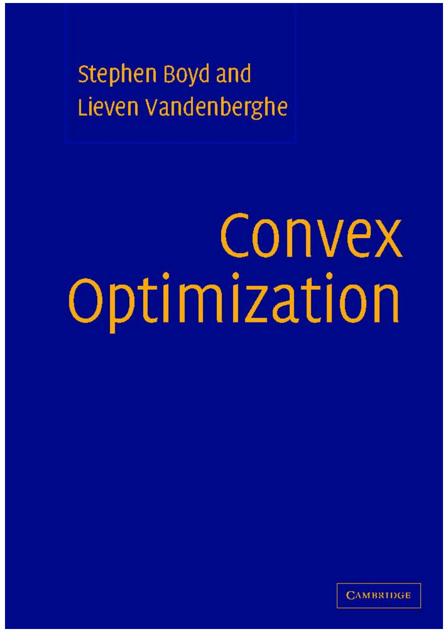
Then

$$\lim_{k \to \infty} \nabla f(\mathbf{x}_k) \to 0$$

with probability 1!

Details in "Online Learning and Stochastic Approximations" by Leon Bottou

2. Solution-Time Guarantees for Convex Optimization Problems



This is not a course on convex optimization, but the field is extremely rich. See *Convex Optimization* by Boyd and Vandenberghe for:

- Rules for proving convexity
- Examples showing how seemingly nonconvex problems can be made convex
- Algorithms for solving convex optimization problems (focus is on interior point)
- Theoretical guarantees on performance

2. Solution-Time Guarantees for Convex Optimization Problems

Most important properties of convex optimization problems:

- Any local minimum is a global minimum
- A global minimum can be found in a predictable, finite number of steps

Example from Boyd and Vandenberghe, section 11.5

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & Ax = b, \end{array}$$

Additional assumptions:

- $t f_0 + \phi$ is *self concordant*
- Sublevel sets are bounded

ϕ is the barrier, e.g. $\phi = -\sum_{i=1}^{m} \log(-f_i)$

self concordant means $|f'''(x)| \leq 2 f''(x)^{3/2}$

2. Solution-Time Guarantees for Convex Optimization Problems

Interior point method:

1. Solve "Phase 1" problem to find a feasible point

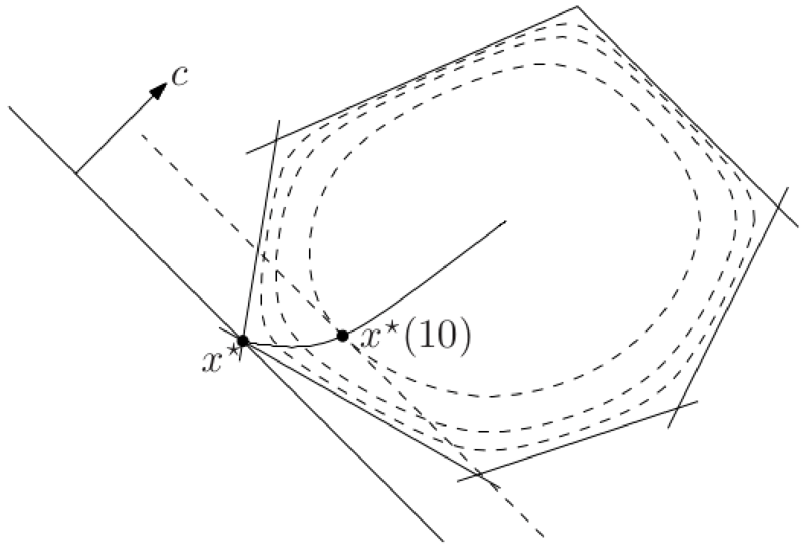
$$\begin{array}{ll} \text{minimize} & s \\ \text{subject to} & f_i(x) \leq s, \quad i = 1, \ldots, m \\ & Ax = b \end{array}$$

2. Start with $t = t^{(0)}$ and Loop

    1. Solve

    $$\begin{array}{ll} \textbf{maximize} & t f_0 + \phi \\ \textbf{subject to} & f_i \leq 0, \quad Ax = b \end{array}$$

    2. $t \leftarrow \mu t$



Solution to each barrier subproblem moves along the "central path"

2. Solution-Time Guarantees for Convex Optimization Problems

Interior point method:

1. Phase I: find a feasible point

$$\begin{array}{ll} \text{minimize} & s \\ \text{subject to} & f_i(x) \le s, \quad i = 1, \ldots, m \\ & Ax = b \end{array}$$

2. Phase II: Start with $t = t^{(0)}$ and Loop:

   1. Solve

$$\begin{array}{ll} \text{maximize} & tf_0 + \phi \\ \text{subject to} & Ax = b \end{array}$$

   2. $t \leftarrow \mu t$

Section 9.6.4 shows that the number of Newton steps for a convex self-concordant problem without inequality constraints is

$$\frac{f(x^{(0)}) - p^\star}{\gamma} + \underbrace{\log_2 \log_2(1/\epsilon)}_{c\,=\,6}$$

$$\frac{1}{\gamma} = \frac{20 - 8\alpha}{\alpha\beta(1-2\alpha)^2}$$ α, β: backtracking params

$p^\star$: optimal value, ϵ: accuracy

For entire Phase II:

$$N = \left\lceil \frac{\log(m/(t^{(0)}\epsilon))}{\log \mu} \right\rceil \left(\frac{m(\mu - 1 - \log \mu)}{\gamma} + c \right)$$

$$c = 6, \qquad \gamma = 1/375, \qquad m/(t^{(0)}\epsilon) = 10^5, \qquad m = 100 \longrightarrow N \approx 8000$$

2. Solution-Time Guarantees for Convex Optimization Problems

Interior point method:

1. Phase I: find a feasible point

$$\begin{array}{ll} \text{minimize} & s \\ \text{subject to} & f_i(x) \le s, \quad i = 1, \ldots, m \\ & Ax = b \end{array}$$

2. Phase II: Start with $t = t^{(0)}$ and Loop:
    1. Solve
    $$\begin{array}{ll} \text{maximize} & t f_0 + \phi \\ \text{subject to} & Ax = b \end{array}$$
    2. $t \leftarrow \mu t$

$$N_{\mathrm{I}} = \left\lceil \sqrt{m+2} \log_2 \frac{(m+1)(m+2)GR}{|\bar{p}^\star|} \right\rceil \left(\frac{1}{2\gamma} + c \right)$$

R: Radius of feasible set

$\bar{p}^\star$: optimal value of Phase I

$$G = \max_i \|\nabla f_i(0)\|_2$$

$$N_{\mathrm{II}} = \left\lceil \sqrt{m+1} \log_2 \frac{(m+1)(M - p^\star)}{\epsilon} \right\rceil \left(\frac{1}{2\gamma} + c \right)$$

$$M \ge \max\left\{ f_0(x^{(0)}), p^\star \right\}$$

Depends on $\sqrt{m}$, $\dim(x)^3$, $\log_2 \frac{GR}{|\bar{p}^\star|}$, $\log_2 \frac{M - p^\star}{\epsilon}$.