FASTER CONVEX OPTIMIZATION

SIMULATED ANNEALING WITH AN EFFICIENT UNIVERSAL BARRIER

JACOB ABERNETHY — UNIVERSITY OF MICHIGAN
(JOINT WORK WITH ELAD HAZAN — PRINCETON)
THIS TALK — OUTLINE

1. The goal of Convex Optimization
2. Interior Point Methods and Path following
3. Hit-and-Run and Simulated Annealing
4. The Annealing-IPM Connection
5. Faster Optimization
GENERAL CONVEX OPTIMIZATION PROBLEM

- Let $K$ be a bounded convex set, we want to solve

$$
\min_{x \in K} \theta^\top x
$$

- Can always convert non-linear objective into a linear one

$$
\min_{x \in K} f(x) \quad \rightarrow \quad \min_{(x,c) \in K \times \mathbb{R}} \quad \begin{cases} 
   c & 
   \text{if } f(x) \leq c 
\end{cases}
$$
THE GRADIENT DESCENT ALGORITHM

- The gradient descent algorithm:

\[
\text{For } t = 1, 2, \ldots :
\]
\[
\tilde{x}_t = x_{t-1} - \eta \nabla f(x_{t-1})
\]
\[
x_t = \text{Proj}_K(\tilde{x}_t)
\]

- Challenge: the Projection step can often be just as hard as the original optimization
BETTER: USE THE CURVATURE — NEWTON’S METHOD

- Newton’s Method is a “smarter” version of gradient descent, moves along the gradient after a transformation

For $t = 1, 2, \ldots$:

$$\tilde{x}_t = x_{t-1} - \nabla^{-2} f(x_{t-1}) \nabla f(x_{t-1})$$

$$x_t = \text{Proj}_K(\tilde{x}_t)$$

GOOD: Typically this step is not required

BAD: Need to invert $N \times N$ mtx requires possible $O(n^{^2.373\ldots})$
For a quadratic function, one only needs a single newton step to reach the global minimum.

For $t = 1, 2, \ldots$:
\[
\tilde{x}_t = x_{t-1} - \nabla^2 f(x_{t-1}) \nabla f(x_{t-1})
\]
\[
x_t = \text{Proj}_K(\tilde{x}_t)
\]
WAIT! OUR ORIGINAL OBJECTIVE ISN’T CURVED . . .

- How does this help us with linear optimization?

$$\min_{x \in K} \theta^\top x + \phi(x)$$

- Add a curved function $\phi()$ to the objective!

- $\phi()$ should be “super-smooth” (more on this later)

- $\phi()$ should be a “barrier”, i.e. goes to $\infty$ on the boundary, but not too quickly!
OPTIMIZATION WITHOUT A BARRIER

\[
\min_{x \in K} \theta^\top x
\]
OPTIMIZATION WITH A BARRIER

$$\min_{x \in K} \theta^\top x + \phi(x)$$
WHAT IS A GOOD BARRIER?

- What is needed for this “barrier func.” \( \phi() \)?

- Canonical example: if set is a polytope \( K = \{x : Ax \leq b\} \) then the logarithmic barrier suffices: \( \phi(x) = -\sum_i \log(b_i - A_i x) \)

- In general, Nesterov and Nemirovski proved that the following two conditions are sufficient. Any function satisfying these conditions is a self-concordant barrier:

\[
\nabla^3 \phi[h, h, h] \leq 2(\nabla^2 \phi[h, h])^{3/2}, \quad \text{and} \\
\nabla \phi[h] \leq \sqrt{\nu \nabla^2 \phi[h, h]},
\]

- \( \nu \) is the barrier parameter which will be important later
ALGORITHM: INTERIOR POINT PATH FOLLOWING METHOD

- Nesterov and Nemirovski developed the sequential “path following” method, described as follows:

  - Let $\alpha = (1 + 1/\sqrt{\nu})$ the “inflation” rate
  - For $t=1,2,...$

1. Update temperature: $f_k(x) := \alpha^k (\theta^\top x) + \phi(x)$

2. Newton update: $\hat{x} \leftarrow \hat{x} - \frac{1}{1+c_k} \nabla^{-2} f_k(\hat{x}) \nabla f_k(\hat{x})$
WHAT DOES THE SEQUENCE OF OBJECTIVES LOOK LIKE?

\[ f_k(x) := \alpha^k (\theta^\top x) + \phi(x) \]

- Let’s show these objective function as we increase \( k \)!!
WHY IS THIS CALLED “PATH FOLLOWING”?

\[ \Phi(\alpha) := \arg \min_{x \in K} \alpha(\theta^\top x) + \phi(x) \]

- As we increase inflation, the minimizer moves closer to the true desired minimum. We can plot this minimizer as \( \alpha \) increases. This is known as the Central Path.
CONVERGENCE RATE OF PATH FOLLOWING

- Nesterov and Nemirovski showed:

  1. Best inflation rate is \( \alpha_k = (1 + 1/\sqrt{\nu})^k \)
  2. Approx error after \( k \) iter is \( \epsilon = \frac{\nu}{(1+1/\sqrt{\nu})^k} \)
  3. Hence, to achieve \( \epsilon \) error, need \( k = O(\sqrt{\nu} \cdot \log(\nu/\epsilon)) \)

- The barrier parameter \( \nu \) is pretty important. Nesterov and Nemirovski showed that every set has a self-concordant barrier with barrier parameter \( \nu = O(n) \)
Define:

\[ \|v\|_x := \sqrt{v^T \nabla^2 \varphi(x) v}, \]  

“local norm” of \( v \) w.r.t. \( x \);

\[ \|v\|_x^* := \sqrt{v^T \nabla^{-2} \varphi(x) v}, \]  
dual norm of \( v \),

\[ \Phi_k(x) := \alpha_k \hat{\theta}^T x + \varphi(x) \]

\[ \lambda(x, \alpha_k) := \|\nabla \Phi_k(x)\|_x^*, \]  
the Newton decrement of \( x \).

**Lemma 1.** Let \( \alpha > 0 \) be arbitrary and let \( \alpha' = \alpha \left(1 + \frac{c}{\sqrt{\nu}}\right) \). Then for any \( x \in \text{int}(K) \), we have \( \lambda(x, \alpha') \leq (1 + c) \lambda(x, \alpha) + c \).

**Lemma 2.** Let \( x \) be arbitrary. Then \( x' \), the newton update from \( x \), satisfies \( \lambda(x', \alpha) \leq 2\lambda^2(x, \alpha) \).
With previous two lemmas, we can prove a very simple invariant.

**Lemma 3.** Let $c = 1/20$. Then for all $k$ we have $\lambda(\hat{x}_k, \alpha_k) < \frac{1}{3}$.

*Proof.* (By induction) The base case is satisfied since we assume that $\lambda(\hat{x}_0, \alpha_0) = 0$, as $\alpha_0 = 1$. Inductive step: assume $\lambda(\hat{x}_{k-1}, \alpha_{k-1}) < 1/3$. Then

$$
\lambda(\hat{x}_k, \alpha_k) \leq 2\lambda^2(\hat{x}_{k-1}, \alpha_k) \\
\leq 2((1 + c)\lambda(\hat{x}_{k-1}, \alpha_{k-1}) + c)^2 < 2(0.4)^2 < 1/3.
$$

The first inequality follows by Lemma 2 and the second by Lemma 1, hence we are done. □
THE PROBLEM: EFFICIENT SELF-CONCORDANT BARRIER IN GENERAL?

- Given any convex set $K$, how can we construct a self-concordant barrier for $K$?
- Polytopes are easy. So are $L2$-balls. We have barriers for some other sets also, e.g. the PSD cone.
- PROBLEM: Find an efficient universal barrier construction?
- Open problem for some time.
THIS TALK — OUTLINE

1. The goal of Convex Optimization
2. Interior Point Methods and Path following
3. *Hit-and-Run and Simulated Annealing*
4. The Annealing-IPM Connection
5. Faster Optimization
SIMULATED ANNEALING FOR OPTIMIZATION

- From Wikipedia: Optimization of a 1-dimensional function
INTRODUCTION TO SIMULATED Annealing

- Your goal is to solve the optimization

\[
\min_{x \in K} f(x)
\]

- Maybe it is easier to *sample from* the distribution

\[
P_t(x) = \frac{\exp(-f(x)/t)}{\int_K \exp(-f(x')/t)dx'}
\]

for a *temperature parameter* $t$
Why is sampling easier? And why would it help anyway?

First, when $t$ is very large, sampling from $P_t()$ is equivalent to sampling from the uniform distribution on $K$. Easy(ish)!

Second, when $t$ is very small, all mass of $P_t()$ is concentrated around minimizer of $f(x)$. That’s what we want!

Third, the successive distributions $P_t()$ and $P_{t+1}()$ are all very close, so we can “warm start” from previous samples.
HIT-AND-RUN FOR LOG-CONCAVE DISTRIBUTIONS

\[ P_t(x) = \frac{\exp(-f(x)/t)}{\int_K \exp(-f(x')/t) dx'} \]

- Notice that \( f() \) convex in \( x \) \( \Rightarrow \) \( \log P_t \) is concave in \( x \)

- Lovasz/Vempala showed that problem of sampling log-concave dists is poly-time using Hit-And-Run random walk
  ........ IF you have a warm start (more on this later)

- Hit-And-Run is an interesting randomization procedure to sample from a convex body, with an interesting history
WHO INVENTED HIT-AND-RUN?


Recent IOE courses taught: 316, 512, 515, 600, 616, 712

Office: 1733 IOE
Phone: (734) 764-6473
Web page: www-personal.umich.edu/~rsmith/

Email address: rsmith@umich.edu

Dr. Smith is the Altarum/ERIM Russell D. O’Neill Professor Emeritus of Engineering and Professor Emeritus of Industrial and Operations Engineering, University of Michigan.
HIT-AND-RUN

Inputs: distribution $P$, #iter $N$, initial $X_0 \in K$.

For $i = 1, 2, \ldots, N$

1. Sample random direction $u \sim N(0, I)$

2. Compute line segment $R = \{X_{i-1} + \rho u : \rho \in \mathbb{R}\} \cap K$

3. Sample $X_i$ from $P$ restricted to $R$

Return $X_N$

Claim: Hit-And-Run walk has stationary distribution $P$

Question: In what way does $K$ enter into this random walk?
HIT-AND-RUN
HIT-AND-RUN REQUIRES ONLY A MEMBERSHIP ORACLE

- Notice: a single update of Hit-And-Run required only computing the endpoints of a line segment.
- Can be accomplished using binary search with a membership oracle
Kalai and Vempala (2006) gave a poly-time guarantee for annealing using Hit-and-Run (membership oracle only!)

1. Sample from $P_k(x) \propto \exp(-\theta^\top x/t_k)$

2. Successive dists are “close enough” if $\text{KL}(P_{k+1}(x) || P_k(x)) \leq 1/2$

3. The closeness is guaranteed as long as $t_k \approx (1 - 1/\sqrt{n})^k$

4. Roughly $O(\sqrt{n \log 1/\epsilon})$ phases needed, $O(n^3)$ Hit-and-Run steps needed for mixing, and $O(n)$ samples needed per phase

- Needed: $n^3$ steps to mix, $n$ samples per phase, $n^{0.5}$ phases. Hence, total running time is $O(n^{4.5})$
FAST CONVEX OPT - SIMULATED ANNEALING - INTERIOR POINT METHODS
FAST CONVEX OPT-SIMULATED ANNEALING-INTERIOR POINT METHODS
FAST CONVEX OPT-SIMULATED ANNEALING-INTERIOR POINT METHODS
We can define a path according to the sequence of means one obtains as we turn down the temperature. Let

\[ \chi(t) := \mathbb{E}_{X \sim \exp(-\theta^T x/t)/Z} [X] \]

be the HeatPath.
TWO DIFFERENT CONVEX OPTIMIZATION TECHNIQUES

Simulated Annealing via Hit-and-Run

Interior Point Methods via Path Following
THE EQUIVALENCE OF THE CENTRAL PATH AND THE HEAT PATH

- Key result of A./Hazan 2015: there exists a barrier function $\phi()$ such that the CentralPath (for $\phi()$) is *identically* the HeatPath for the sequence of annealing distributions.

These are the same object.
WHAT IS THE SPECIAL BARRIER?

- The barrier $\phi()$ corresponds to the “differential entropy” of the exponential family distribution. Equivalently, it’s the Fenchel conjugate of the log-partition function.

  - Let $A(\theta) = \log \int_K \exp(\theta^\top x) dx$
  - Let $A^*(x) = \sup_\theta \theta^\top x - A(\theta)$
  - A fact about exponential families: $\nabla A(\theta) = \mathbb{E}_{X \sim P_\theta}[X]$
  - A fact about Fenchel duality: $\nabla A(\theta) = \arg \max_{x \in K} \theta^\top x - A^*(x)$

- Guler 1996 showed this function is a barrier for cones. Bubeck and Eldan 2015 showed this in general, and gave an optimal parameter bound of $n(1 + o(1))$. 
**SIMULATED ANNEALING <=> IPM PATH FOLLOWING**

Define: \[
\left\| \frac{\mu}{\pi} \right\| \equiv \mathbb{E}_{x \sim \mu} \left( \frac{\mu(x)}{\pi(x)} \right) = \int_{x \sim \mu} \left( \frac{\mu(x)}{\pi(x)} \right) d\mu(x).
\]

Let \( \gamma > 0 \) be a parameter

**Lovasz/Vempala:** Properties of log-concave distributions imply that, given a warm start from \( P_\theta \), Hit-and-Run mixes quickly (in \( \tilde{O}(n^3) \) steps) to \( P_{(1+\gamma)\theta} \) as long as \( \left\| \frac{P_{(1+\gamma)\theta}}{P_\theta} \right\| = O(1) \)

**Nesterov/Nemirovski:** Properties of self-concordant barrier functions imply that, for the iterative path following scheme, a single newton step suffices to maintain the invariant \( \lambda(x_k, (1 + \gamma)^k) < 1/3 \).
Question: How to bound $\left\| \frac{P_{(1+\gamma)\theta}}{P_\theta} \right\|$ using IPM analysis?

Consider the log of the L2-distance between distributions:

$$\log \left\| \frac{P_{(1+\gamma)\theta}}{P_\theta} \right\| = KL(P_{(1+\gamma)\theta} \| P_\theta) + KL(P_{(1-\gamma)\theta} \| P_\theta)$$

$$= D_A((1 + \gamma)\theta, \theta) + D_A((1 - \gamma)\theta, \theta)$$

$$\approx 2\lambda^2(x((1 + \gamma)\theta), 1)$$

exact quantity of interest in IPM
CONCLUSIONS

- **Observation**: Interior Point Path Following surprisingly equivalent to Simulated Annealing

- **Benefit 1**: This unifies two rich research areas, and lets one borrow tricks from barrier methods to understand annealing, and vice versa

- **Benefit 2**: The connection allows us to get a speedup on annealing using barrier methods, improving Kalai/Vempala’s rate of $O(n^{4.5})$ to $O(n^{1/2}n^4)$
SOME COMMENTS ON CONNECTIONS TO UNIVERSAL BARRIER

- jake use the whiteboard...
FIN