Escaping the Local Minima via Simulated Annealing: Optimization of Approximately Convex Functions

Tengyuan Liang

Department of Statistics, The Wharton School
University of Pennsylvania

Joint work with Alex Belloni, Hari Narayanan and Sasha Rakhlin

Problem formulation

$\mathcal{K} \subset \mathbb{R}^n$ convex bounded set with membership or separation oracle.

Unknown convex Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$.

Noisy access to function values (stochastic zeroth order oracle):

$$\texttt{Query } x \in \mathcal{K}\texttt{, observe } y = f(x) + \eta.$$

where $\mathbb{E}[\eta] = 0$, say sub-gaussian.

What is the oracle complexity of finding ϵ-minimizer?

Answer: $\mathcal{O}^*(n^{\alpha}\epsilon^{-2})$

(Shamir '15): the lower bound for linear f is $\mathcal{O}(n^2\epsilon^{-2})$

Why?

Examples:

- Two-stage stochastic programming

$$\max_{x \in \mathcal{K}} \quad p \cdot x + \mathbb{E}_\eta \left[\max \{ q \cdot y \mid Ay \leq Bx - \eta \} \right]$$

- Privacy-preserving regression

$$\min_{w \in \mathcal{K}} \quad \frac{1}{n} \sum_{i=1}^{n} (x_i \cdot w - y_i)^2$$

- Online prediction

$$\min_{x \in \mathcal{K}} \max_{y} \quad \ell(x, y) + \mathbb{E}_\eta \Phi(y, \eta)$$

What stochastic optimization approaches are available?

For *stochastic* zero-order problems, one may:

I. Estimate first-order (gradient) information and proceed with a noiseless first-order method

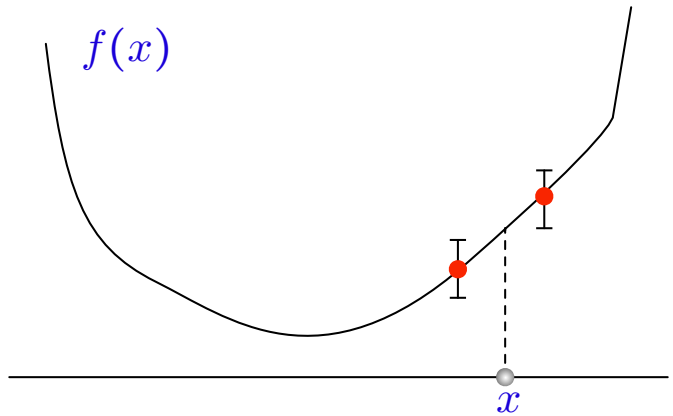
or

II. Estimate zeroth-order information and proceed with a noiseless zeroth order method

or

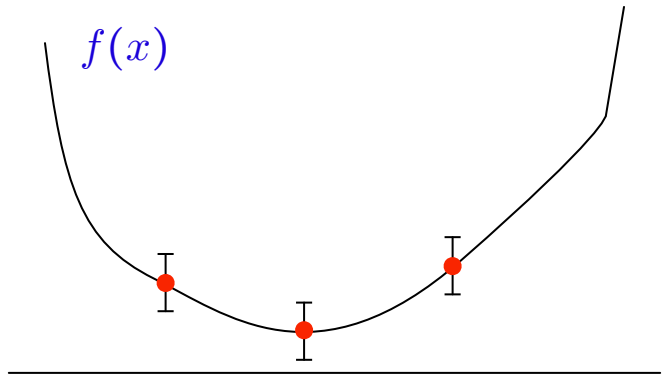
III. Do a random walk and hope it is robust to noise.

Approach I: some of the difficulties



Estimate the gradient when $\mathtt{f}'(\mathtt{x})$ is *small*: oracle complexity dependence on ϵ are worse than ϵ^{-2}.

Approach II: exploit convexity efficiently



Noiseless case in $\mathbb{R}^n$:
"Pyramid" construction of (Nemirovskii & Yudin '79): $\mathcal{O}^*(n^7)$
"Regular T-gons" of (Protasov '96): $\mathcal{O}^*(n^2)$

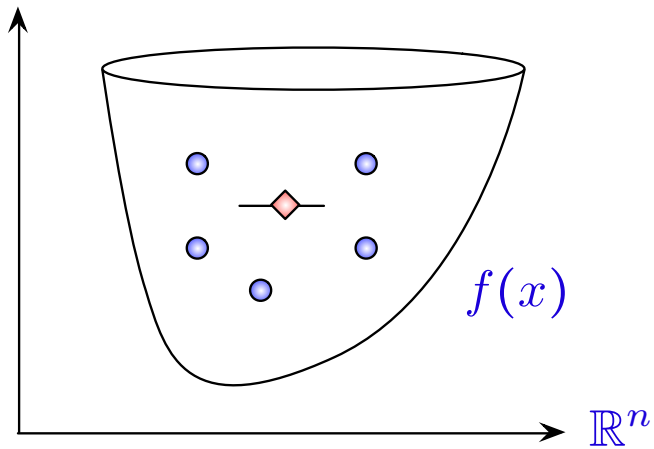
(Agarwal, Foster, Hsu, Kakade, Rakhlin '13): Extend to noisy oracles.
Queries required for *regret*:

$$\mathcal{O}(n^{33}\epsilon^{-2})$$

Is there another zeroth order method for exact oracle that is *robust*?

Approach III, 1st try: random walk on the epigraph

(Bertsimas & Vempala '04):



(L., Narayanan, Rakhlin '14): For noisy access to the set, number of queries is

$$\mathcal{O}^*(n^{14}\epsilon^{-2})$$

Can we do better?

Our Approach

Noisy oracle to approximately convex oracle

- Think of noisy evaluations $f(x) + \eta$ as noiseless answers from a nearly-convex function $F(x)$.
- By querying τ times at x and averaging,

$$|F(x) - f(x)| \leq c/\sqrt{\tau}$$

  with high probability.
- Assuming we do not query at x again, F is "well-defined"

More general problem: approximately convex oracles

Consider an *approximately convex* function $F : \mathbb{R}^n \to \mathbb{R}$: exists convex Lipschitz f s.t.

$$\sup_{\mathcal{K}} |f - F| \leq \epsilon/n$$

`Query` $x \in \mathcal{K}$, `observe` $F(x)$.

What is the oracle complexity of finding ϵ-minimizer for F (and thus f)?

Answer: $\mathcal{O}(n^{\beta} \log(1/\epsilon)^{c})$

Convex function f:

- Sample from log-concave measure $e^{-f(x)/T}$ via Grid-Walk, Ball-Walk or Hit-and-Run, (Applegate, Kannan '91), (Kalai & Vempala '05), (Lovász & Vempala '06).
- Simulated annealing with decreasing temperature, (Kalai & Vempala '05).

Approximately convex function F: our approach

- ? Sampling: *approximately-log-concave* measure $e^{-F(x)/T}$ via geometric random walk
- ? Annealing: optimization of *approximate convex* function F

Oracle complexity

Theorem.

Given that $|F - f| < \mathcal{O}(\epsilon/n)$ (F approximately convex, f convex Lipschitz),

n^3 queries for 1 point $\times$ n parallel strands $\times$ $\sqrt{n}$ epochs $=$ $n^{4.5}$

query complexity are enough to provide an ϵ minimizer to F.

(same as Lovász-Vempala in exact log-concave case)

- Dependence on Lipschitz constant of f, and ϵ is only logarithmic.
- n^1 can be parallelized.
- Can be applied to case with decreasing non-convexity near the optimum.

Simulated annealing

Extension of (Kalai & Vempala '05) to approximate convex F:

◇ Concentration around optimum for low temperature.

Given F is such that $|F - f|_\infty \le \epsilon$, for $X \sim \exp\{-F(x)/T\}$,

$$\mathbb{E}_F f(X) - \min_{x \in \mathcal{K}} f(x) \le (n+1)T \cdot \exp(2\epsilon/T)$$

So, if final temperature $T \propto \epsilon$, we have $\mathcal{O}(n\epsilon)$-minimizer.

◇ Decrease temperature slowly between epochs: warm start.

Let $\pi_{F_i} \propto \exp\{-F(x)/T_i\}$. Choose $T_i = T_{i-1}\left(1 - \frac{1}{\sqrt{n}}\right)$,

$$\|\pi_{F_i}/\pi_{F_{i+1}}\| \le 5\exp(2\epsilon/T_i)$$

We need the final temperature $T_K \propto \epsilon$. Number of epochs K: $\sqrt{n}\log(1/\epsilon)$.

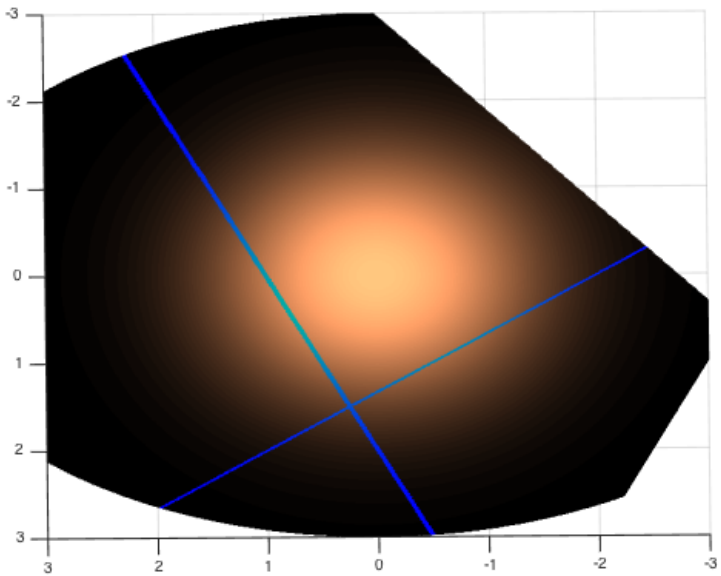
Reshaping/Rounding

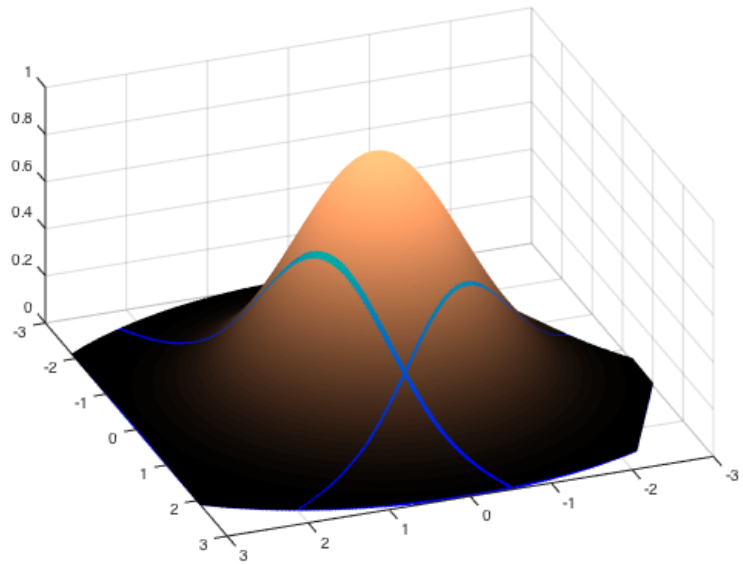
(Guédon and Rudelson '07), (Vershynin '10): Need to run

$$m = n \log n$$

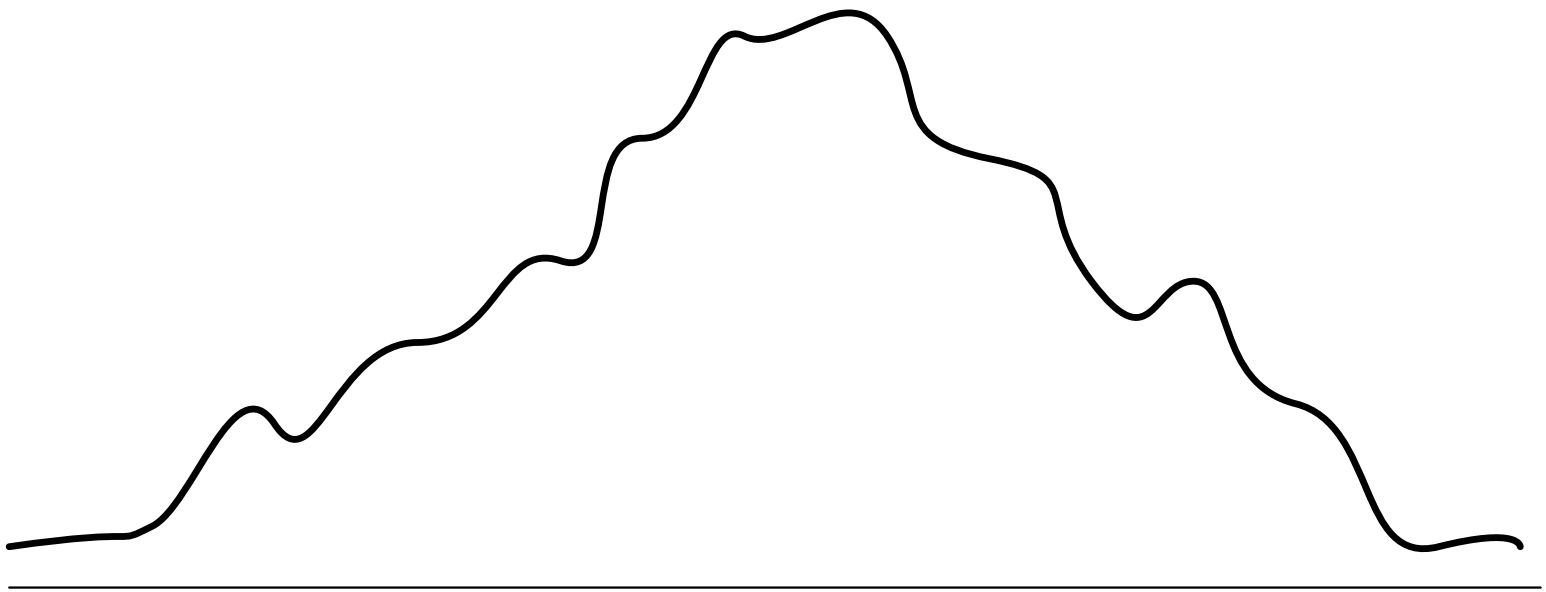
strands of random walks to bring distribution to near-isotropic shape.

Hit-and-Run

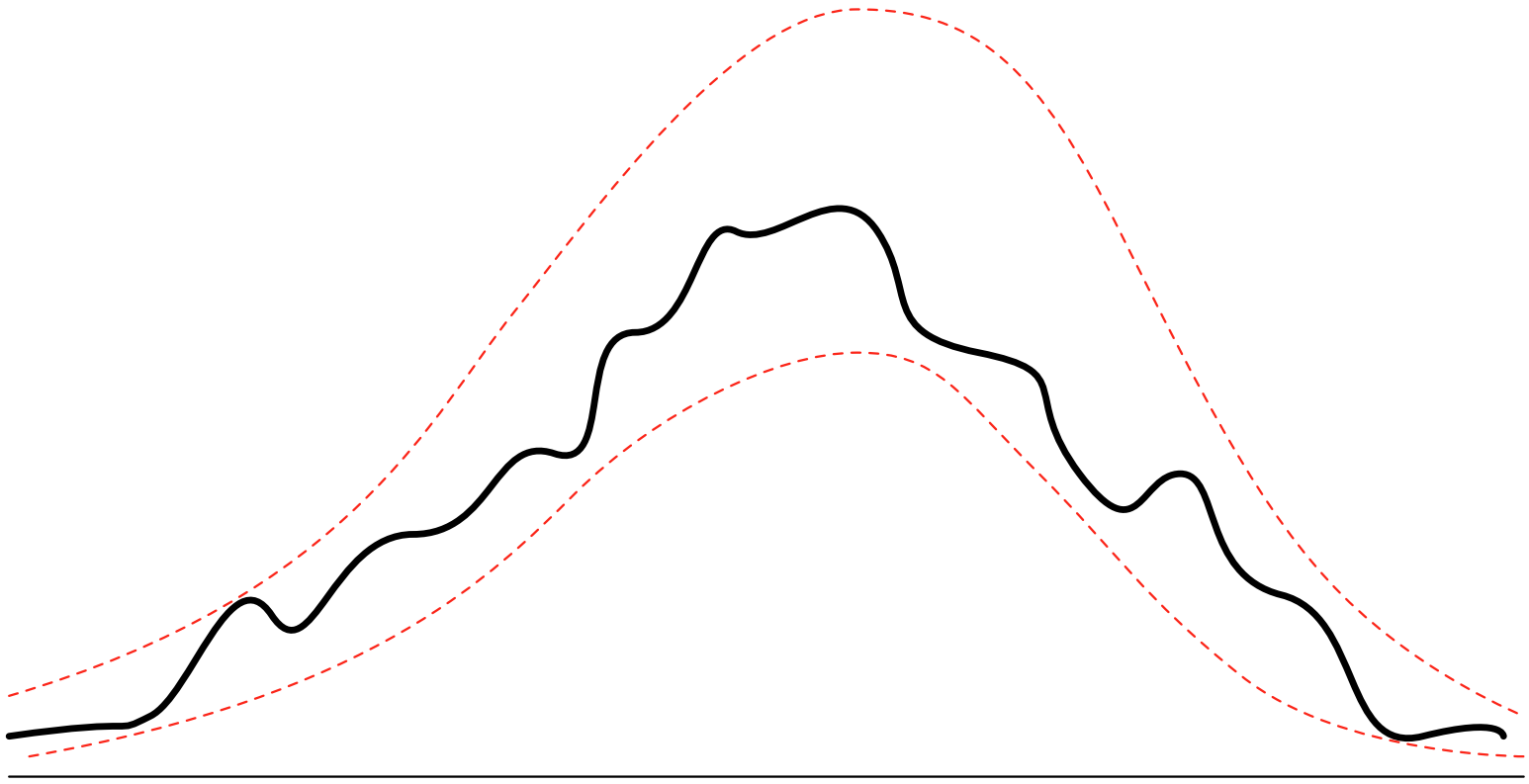




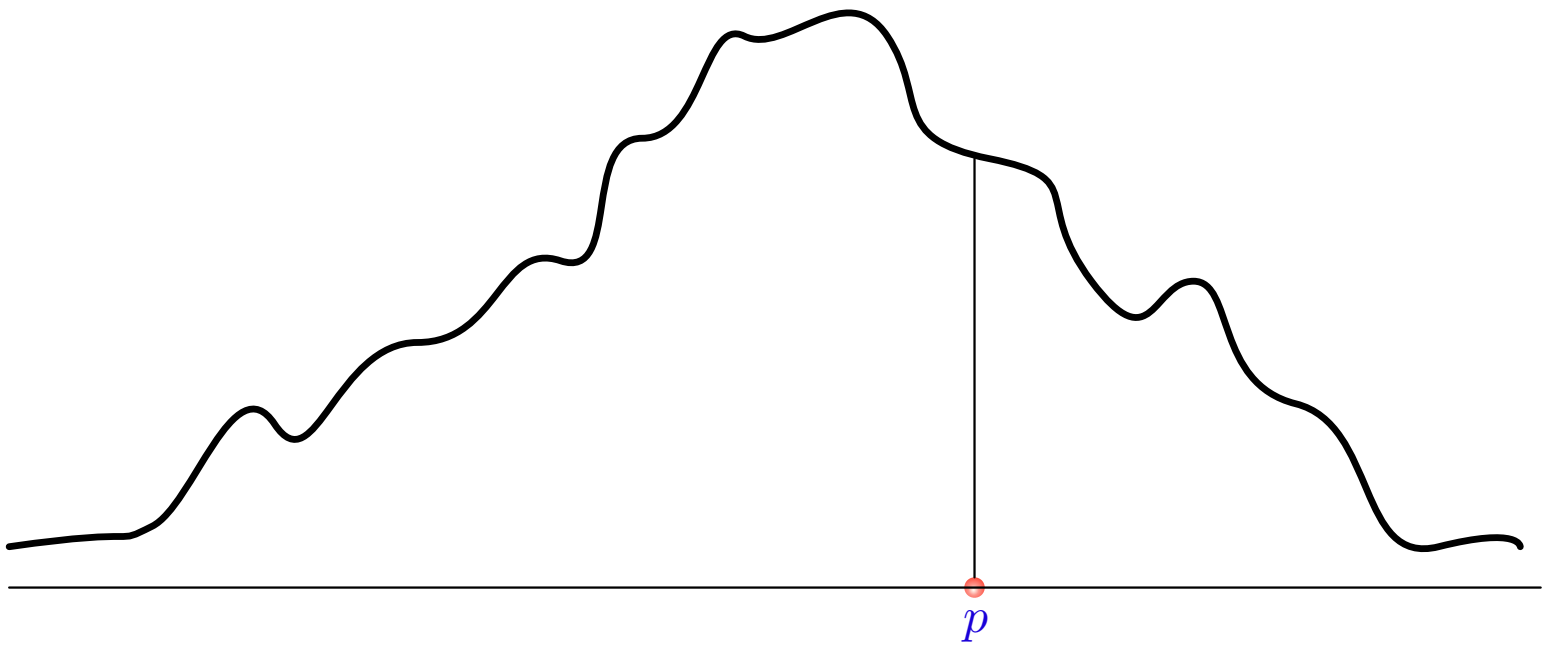
Sample 1-dim nearly log-concave measure



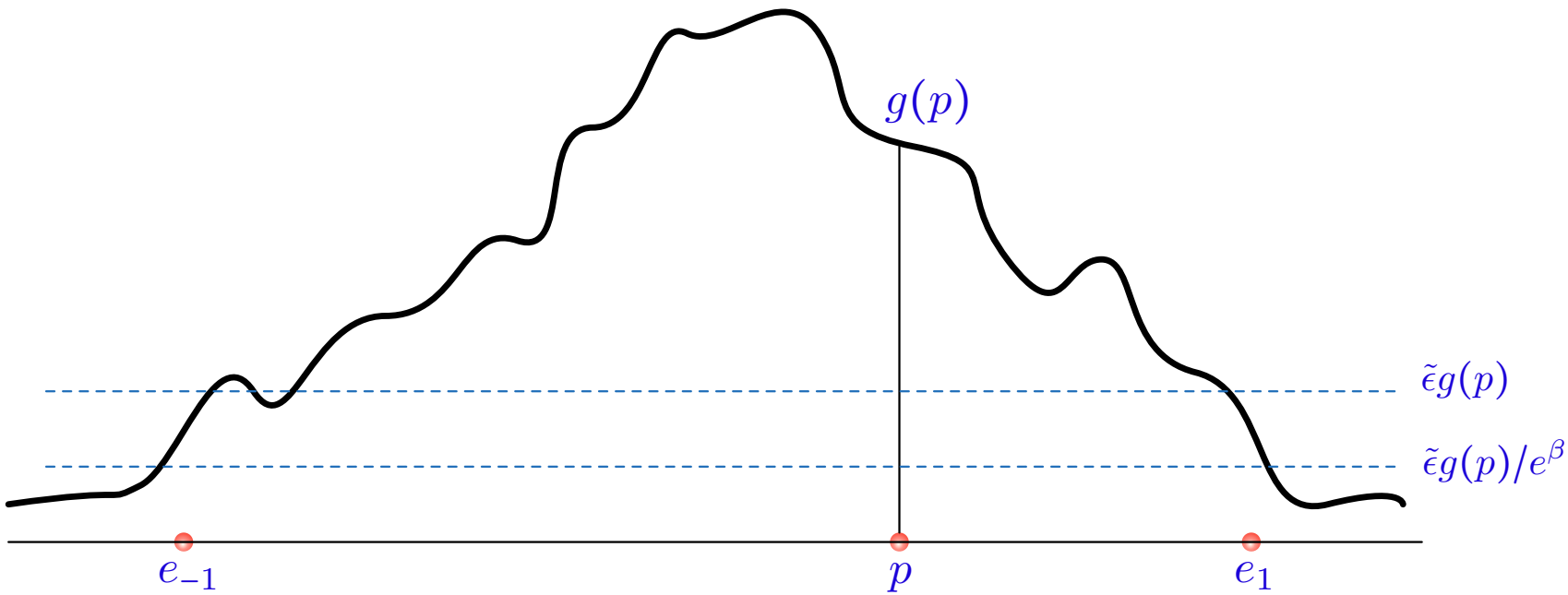
Sample 1-dim nearly log-concave measure



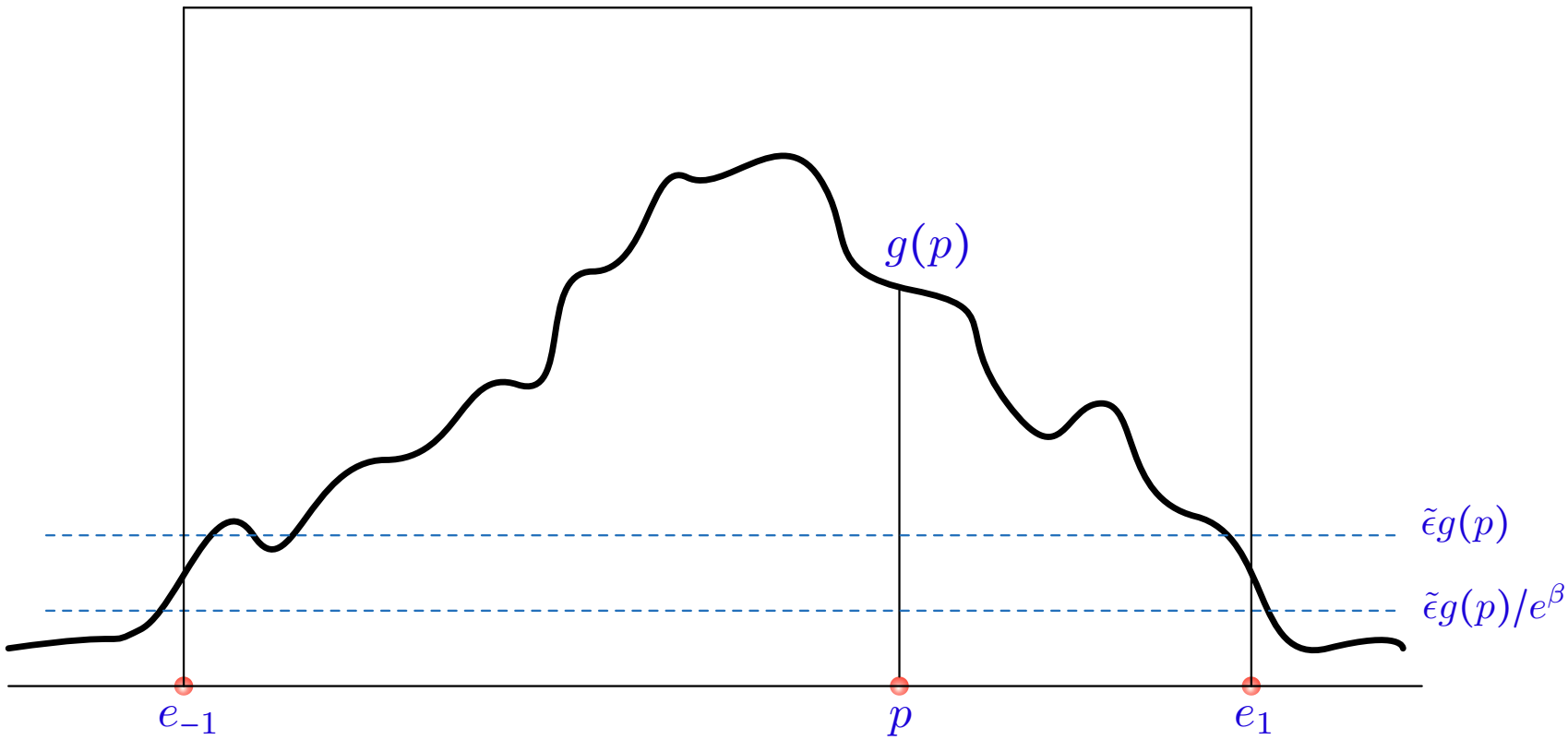
Sample 1-dim nearly log-concave measure



Sample 1-dim nearly log-concave measure



Sample 1-dim nearly log-concave measure

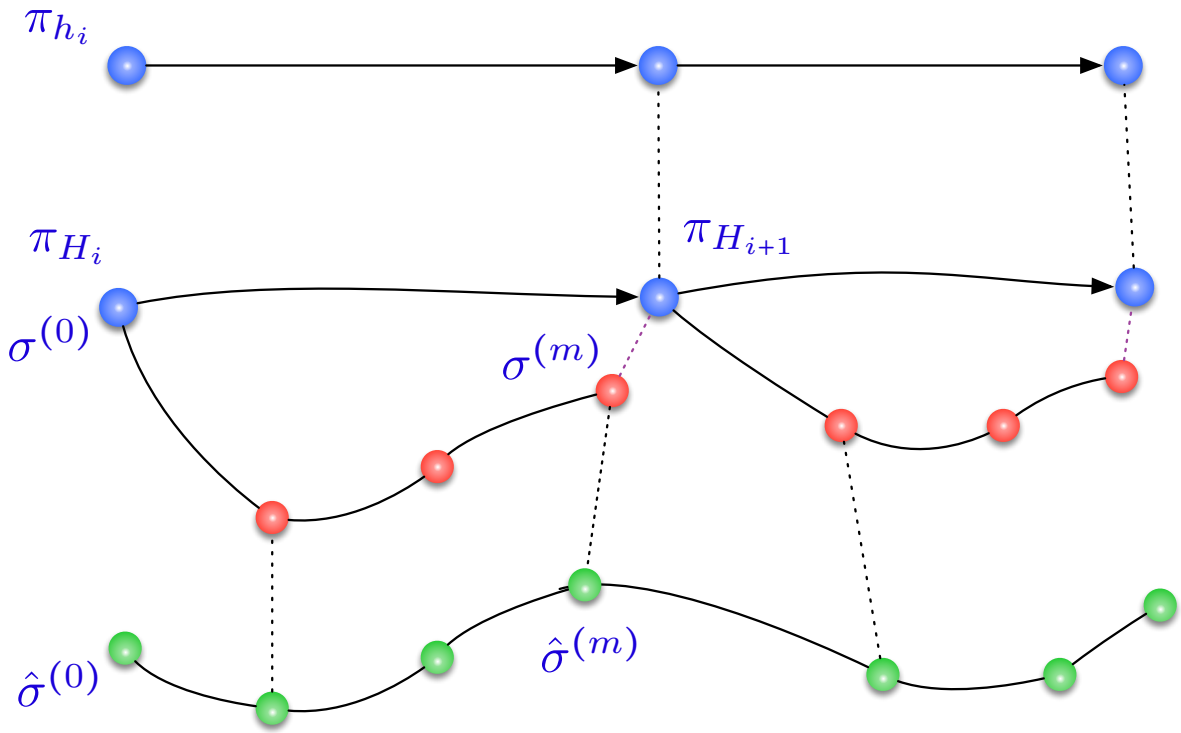


Lemma.

Acceptance probability is at least $\frac{e^{-5\beta}\log 2}{2\log(2/\tilde{\epsilon})}$.

Evolution of distributions

$\pi_{h_i} \propto \exp\{-f(x)/T_i\}$ and $\pi_{H_i} \propto \exp\{-F(x)/T_i\}$



Lower bound on conductance

Let P_x^g denote the transition kernel. Define s-conductance

$$\phi_s^g = \inf_{S \subset \mathcal{K}, s < \pi_g(S) \le 1/2} \frac{\int_{x \in S} P_x^g(\mathcal{K} \backslash S) d\pi_g}{\pi_g(S) - s}.$$

Lemma.

Let g be a $\beta/2$-approximate log-concave measure

$$\exists \quad \text{log-concave} \quad h \quad \text{s.t.} \quad \sup_{\mathcal{K}} |\log g - \log h| \le \beta/2.$$

Then conductance and s-conductance of the random walk induced by g are lower bounded as

$$\phi^g \ge e^{-3\beta} \phi^h \quad \text{and} \quad \phi_s^g \ge e^{-3\beta} \phi_{s/e^\beta}^h.$$

Fast mixing

(Lovász & Simonovits '93):

$$d_{tv}(\pi_g, \sigma^{(m)}) \leq H_s + \frac{H_s}{s}\left(1 - \frac{(\phi_s^g)^2}{2}\right)^m$$

where $H_s = \sup\{|\pi_g(A) - \sigma^{(0)}(A)| : \pi_g(A) \leq s\}$.

Fast mixing

(Lovász & Simonovits '93):

$$d_{tv}(\pi_g, \sigma^{(m)}) \le H_s + \frac{H_s}{s}\left(1 - \frac{(\phi_s^g)^2}{2}\right)^m$$

where $H_s = \sup\{|\pi_g(A) - \sigma^{(0)}(A)| : \pi_g(A) \le s\}$.

Theorem (Mixing Time for Nearly-Log-Concave Measure)**.**

For measure π_H being β-approximately log-concave, let $M = \|\sigma^{(0)}/\pi_H\| = \int (d\sigma^{(0)}/d\pi_H)d\sigma^{(0)}$. Fix $\gamma > 0$. Then

$$m \ge n^2 \frac{R^2}{r^2} C e^{6\beta} \log^2 \frac{e^\beta M n R}{r\gamma^2} \log \frac{M}{\gamma}$$

steps of Hit-and-Run yield

$$d_{tv}(\pi_H, \sigma^{(m)}) \le \gamma.$$

In the paper: guarantee for $\hat{\sigma}^{(m)}$ propagated from actual algorithm. (approximation, truncation, etc.)

Oracle complexity

Given that $|F - f| < \mathcal{O}(\epsilon/n)$

$$n^3 \text{ queries for 1 point} \times n \text{ parallel strands} \times \sqrt{n} \text{ epochs} = n^{4.5}$$

(same as Lovász-Vempala in exact log-concave case)

Oracle complexity

Given that $|F - f| < \mathcal{O}(\epsilon/n)$

$$n^3 \text{ queries for 1 point} \times n \text{ parallel strands} \times \sqrt{n} \text{ epochs} = n^{4.5}$$

(same as Lovász-Vempala in exact log-concave case)

Back to *Stochastic Oracle*:

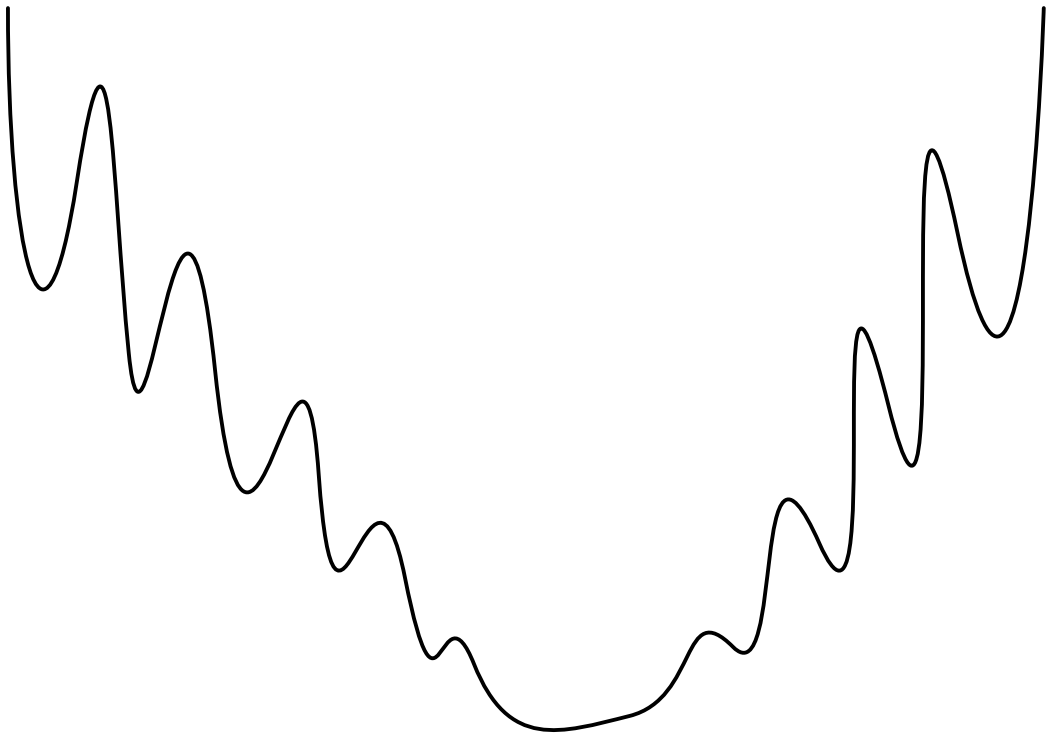
- To attain ϵ/n-accuracy, need $n^2\epsilon^{-2}$ queries per single point to decrease noise level.
- Another factor of n for uniform control of noise over an ϵ-grid.

Total:

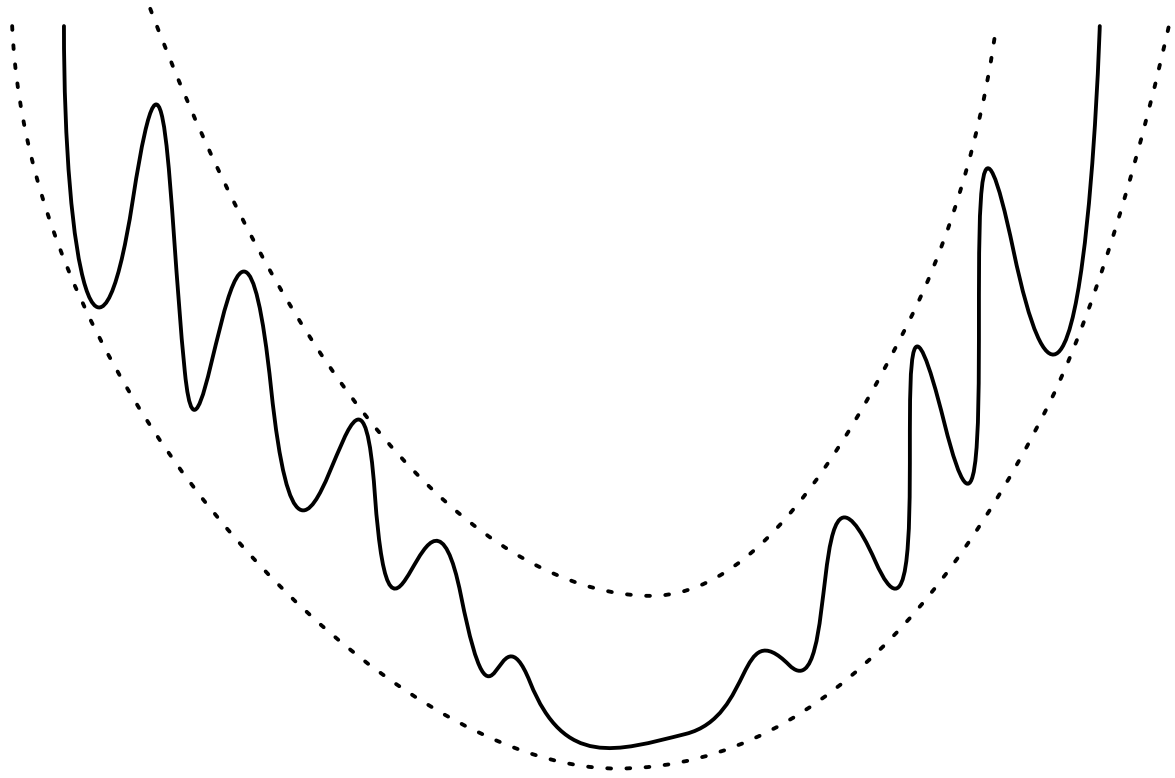
$$n^{7.5}\epsilon^{-2}$$

- Dependence on Lipschitz constant of f is only logarithmic.
- n^1 can be parallelized
- n^1 can be removed if noise has spacial structure

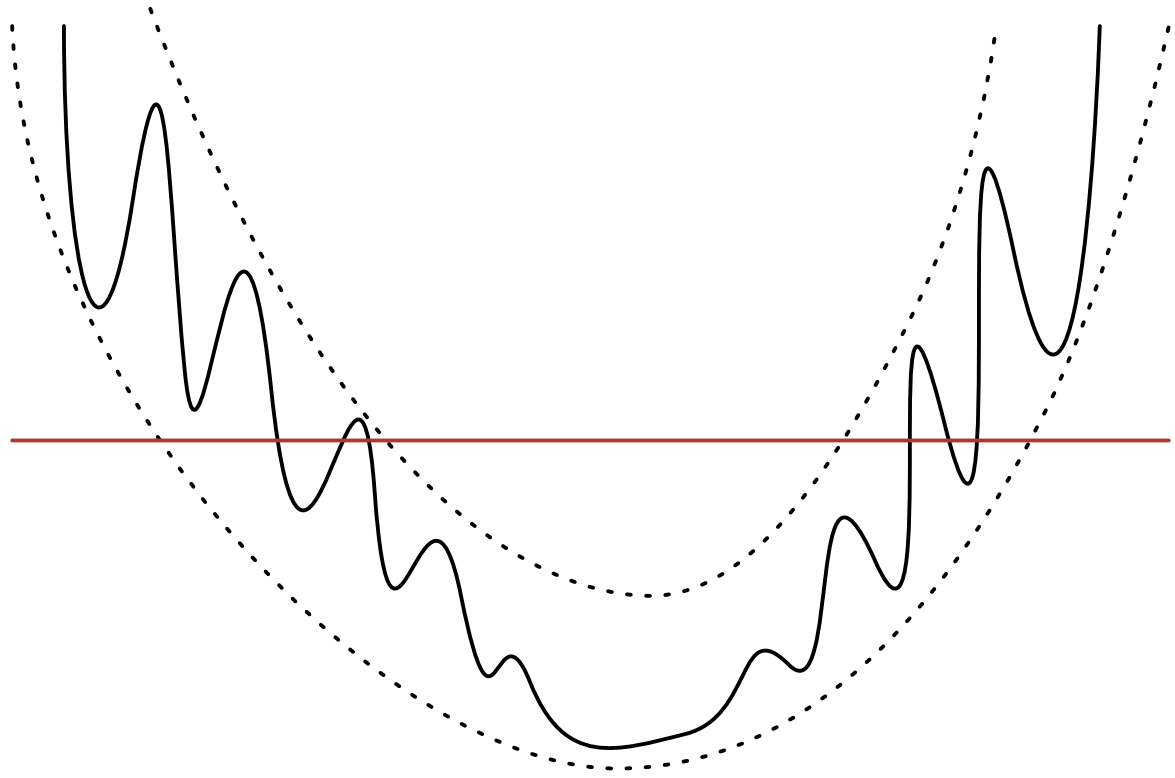
Escaping the local minima



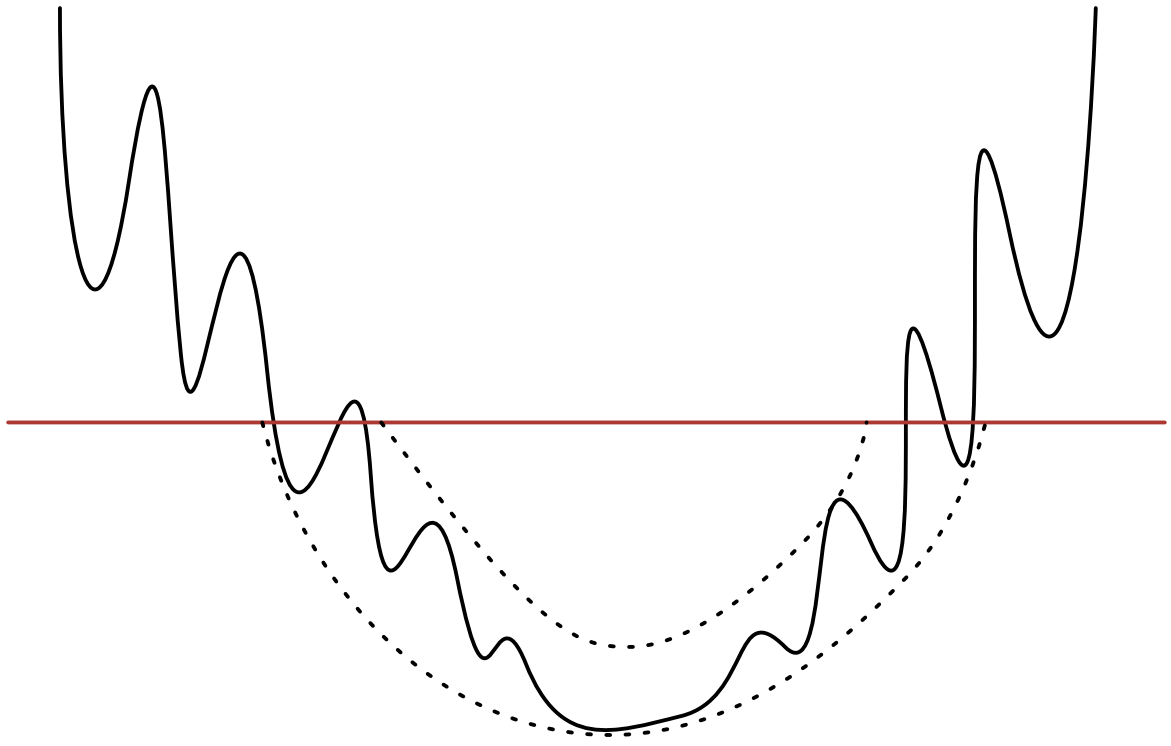
Escaping the local minima



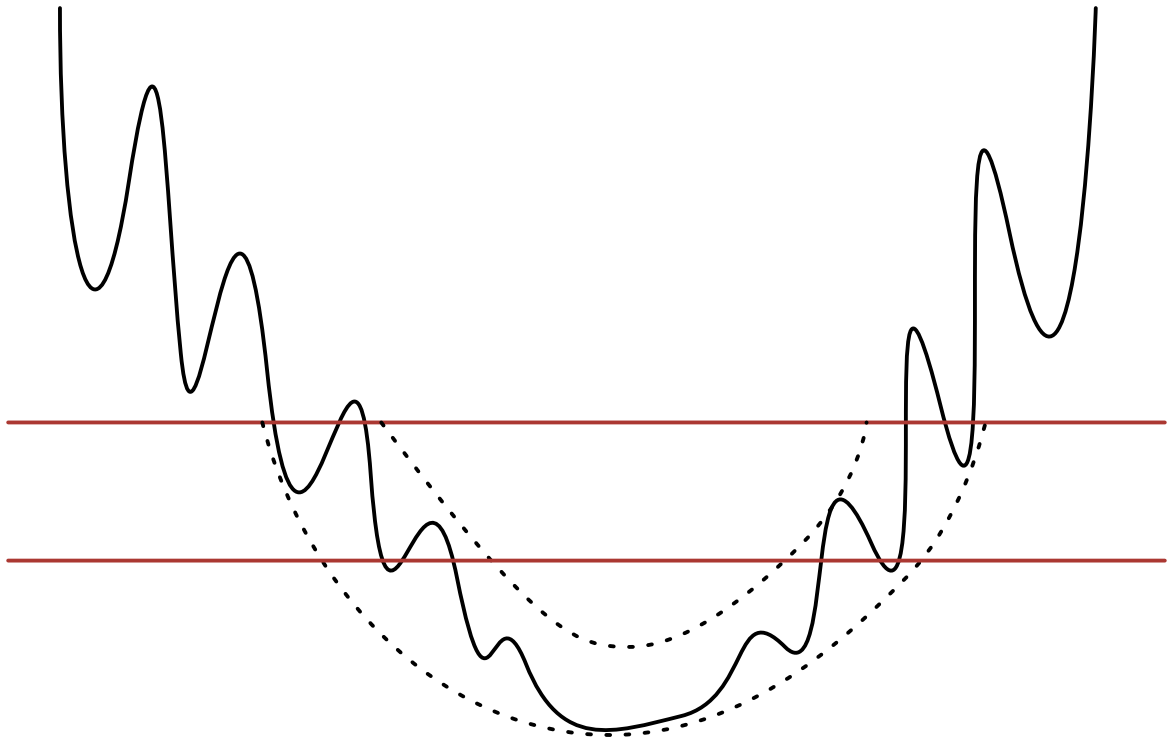
Escaping the local minima



Escaping the local minima



Escaping the local minima



Better query complexity under additional assumptions

Smooth functions:

(Dyer, Kannan, Stougie, '14): *"A simple randomised algorithm for convex optimisation. Application to two-stage stochastic programming"*

Very cute method, no need for simulated annealing. Drawbacks:

- gets stuck for non-smooth functions
- query complexity has low dependence on n, but worse dependence on ϵ

Multiple papers on two-query complexity (Nesterov, Agarwal-Dekel-Xiao, Jamieson-Nowak-Recht, Duchi-Jordan-Wainwright-Wibisono)

Thanks!