STORM: STochastic Optimization using Random Models

Matt Menickelly

Joint work with Ruobing Chen and Katya Scheinberg

Lehigh University

January 5, 2016

The General Problem - Black Box Stochastic Optimization

Want to minimize (unconstrained) $f(x) : \mathbb{R}^n \to \mathbb{R}$. Minimal assumptions: $f \in C^1$ or $f \in C^2$, f is bounded below.

The General Problem - Black Box Stochastic Optimization

Want to minimize (unconstrained) $f(x) : \mathbb{R}^n \to \mathbb{R}$. Minimal assumptions: $f \in C^1$ or $f \in C^2$, f is bounded below.

However, we cannot compute $f(x)$ exactly: only have access to *estimators* $\tilde{f}(x, \omega)$, where $\omega \in \Omega$ is a random variable beyond optimizer's control.

The General Problem - Black Box Stochastic Optimization

Want to minimize (unconstrained) $f(x) : \mathbb{R}^n \to \mathbb{R}$. Minimal assumptions: $f \in C^1$ or $f \in C^2$, f is bounded below.

However, we cannot compute $f(x)$ exactly: only have access to *estimators* $\tilde{f}(x, \omega)$, where $\omega \in \Omega$ is a random variable beyond optimizer's control. This also implies one cannot compute $\nabla f(x)$ or $\nabla^2 f(x)$ exactly - only *estimators* $g(x, \omega)$ or $H(x, \omega)$. Examples to follow immediately.

Gradient Estimators: Supervised Learning/SGD

- Suppose feature-label pairs $(x, y) \in X \times Y \subset \mathbb{R}^n \times \{-1, 1\}$ come from some unknown distribution on $X \times Y$.
- Suppose you have a *training set* of finite size p, $(x^1, y^1), (x^2, y^2), \ldots, (x^p, y^p) \subset X \times Y$.

Task: Letting $\ell(f, x, y)$ denote a *loss* incurred by using $f(x)$ to predict y, minimize $\mathcal{L}(f) = \mathbb{E}_{(x,y)}\ell(f, x, y)$.

Gradient Estimators: Supervised Learning/SGD

- Suppose feature-label pairs $(x, y) \in X \times Y \subset \mathbb{R}^n \times \{-1, 1\}$ come from some unknown distribution on $X \times Y$.
- Suppose you have a *training set* of finite size p, $(x^1, y^1), (x^2, y^2), \ldots, (x^p, y^p) \subset X \times Y$.

Task: Letting $\ell(f, x, y)$ denote a *loss* incurred by using $f(x)$ to predict y, minimize $\mathcal{L}(f) = \mathbb{E}_{(x,y)}\ell(f, x, y)$.

What one does: Let $w \in \mathbb{R}^d$ parameterize a class of functions and approximate $\mathcal{L}(f)$ by

$$\mathcal{L}_p(w) = \frac{1}{p}\sum_{i=1}^{p} \ell(w, x^i, y^i).$$

Gradient Estimators: Supervised Learning/SGD

- Suppose feature-label pairs $(x, y) \in X \times Y \subset \mathbb{R}^n \times \{-1, 1\}$ come from some unknown distribution on $X \times Y$.
- Suppose you have a *training set* of finite size p, $(x^1, y^1), (x^2, y^2), \ldots, (x^p, y^p) \subset X \times Y$.

Task: Letting $\ell(f, x, y)$ denote a *loss* incurred by using $f(x)$ to predict y, minimize $\mathcal{L}(f) = \mathbb{E}_{(x,y)}\ell(f, x, y)$.

What one does: Let $w \in \mathbb{R}^d$ parameterize a class of functions and approximate $\mathcal{L}(f)$ by

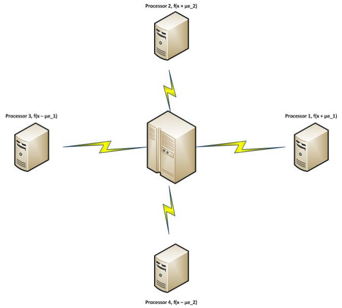
$$\mathcal{L}_p(w) = \frac{1}{p}\sum_{i=1}^{p}\ell(w, x^i, y^i).$$

If $\ell(w, x^i, y^i)$ is smooth, $|\mathcal{S}| \leq p$, then a *gradient estimator* for $\nabla\mathcal{L}(f)$ is

$$g(w) = \frac{1}{|\mathcal{S}|}\sum_{i \in \mathcal{S}}\nabla\ell(w, x^i, y^i).$$

Gradient Estimators: Simulation Optimization

Suppose an unconstrained objective $f(x)$ is approximated via a stochastic simulation $\tilde{f}(x)$.
To *estimate* the gradient $\nabla f(x)$ via central difference gradient, choose a parameter $\mu > 0$, and run the simulation in parallel at "compass points":



Central difference gradient:

$$g(x) = \frac{1}{2\mu} \begin{bmatrix} \tilde{f}(x + \mu e_1) - \tilde{f}(x - \mu e_1) \\ \vdots \\ \tilde{f}(x + \mu e_n) - \tilde{f}(x - \mu e_n) \end{bmatrix}$$

(Kiefer-Wolfowitz)

What Already Exists?

OK, so gradient (and Hessian) approximation aren't abstruse things. Supposing one has access to these things, what already exists?

Stochastic optimization (SO) is a huge field. Arguably the two largest families to solve our problem: stochastic gradient (SG) methods and sample average approximation (SAA) methods.

Stochastic Gradient (SG) Methods

Suppose access to estimator $g(x, \omega)$ of $\nabla f(x)$.

Algorithm 1 Stochastic Gradient Descent (Robbins Monro)

1: Initialize x^0.
2: **while** TRUE **do**
3: $\quad x^{k+1} \leftarrow x^k - \alpha_k g(x^k, \omega_k)$
4: $\quad k \leftarrow k + 1$
5: **end while**

- If $\mathbb{E}_\omega[g(x, \omega)] = \nabla f(x)$ for all x in the search space, then converges in expectation ($\mathbb{E}[f(x^k) - f^*] = \mathcal{O}(1/k)$ in the strongly convex case).
- Need $\alpha_k \to 0$ and $\displaystyle\sum_{k=0}^{\infty} \alpha_k = \infty$.
- Practical performance *heavily* depends on good tuning of $\{\alpha_k\}$.

Sample Average Approximation (SAA) Methods

General flavor: suppose access to unbiased estimators $g(x, \omega)$ of $\nabla f(x)$ and $\tilde{f}(x, \omega)$ of $f(x)$.

In the kth iteration of your favorite iterative algorithm for unconstrained optimization, define a sample size N_k and

$$f_{N_k}(x^k) = \frac{1}{N_k} \sum_{i=1}^{N_k} \tilde{f}(x^k, \omega_i) \quad \nabla_{N_k} f(x^k) = \frac{1}{N_k} \sum_{i=1}^{N_k} g(x^k, \omega_i)$$

- Variants exist that work quite well in practice
- Generally, $\{N_k\}_{k=0}^{\infty}$ must be nondecreasing (variance reduction)
- Strong assumptions necessary for analysis.

Compare & Contrast

(SG)

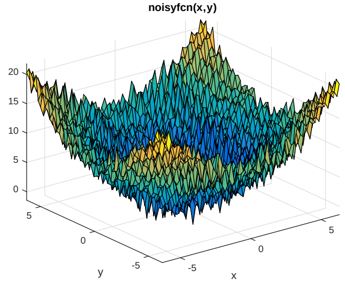
1. Accuracy of $g(x^k, \omega_k)$ does not improve with k
2. Constantly cheap iterations
3. Particular step size restrictions - inflexible
4. Asymptotically optimal rates known
5. INHERENTLY ASSUMES UNBIASED ESTIMATORS

(SAA)

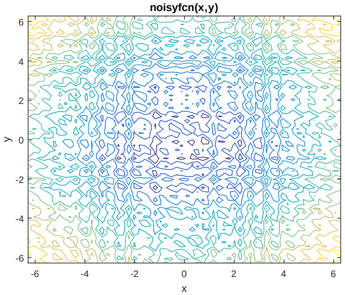
1. Accuracy of $f_{N_k}(x^k)$, $\nabla_{N_k} f(x^k)$ improves with k
2. Iteration complexity grows with N_k
3. Works in many algorithmic frameworks
4. Through adaptive N_k, same optimal rates
5. INHERENTLY ASSUMES UNBIASED ESTIMATORS

STORM

Our method: STORM (STochastic Optimization using Random Models).

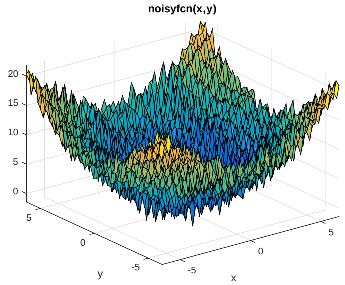


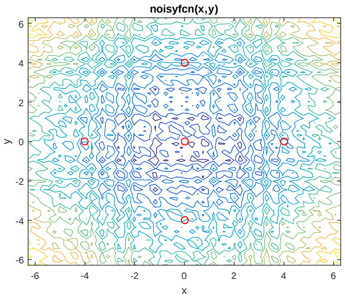
noisyfcn(x,y)
20
15
10
5
0
5
0
-5
y
-5
0
5
x



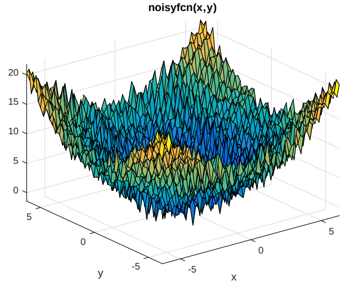
noisyfcn(x,y)
6
4
2
0
-2
-4
-6
y
-6
-4
-2
0
2
4
6
x

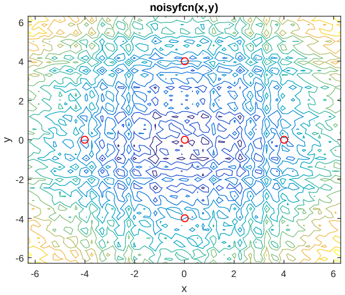
Derivative-Free Optimization - A Brief Intro

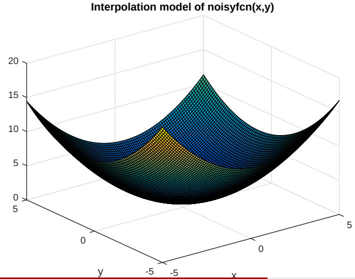


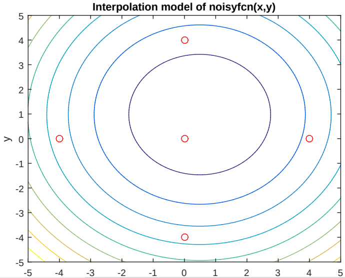


Derivative-Free Optimization - A Brief Intro

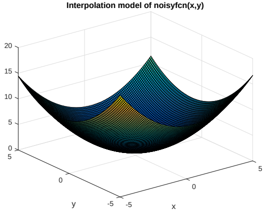


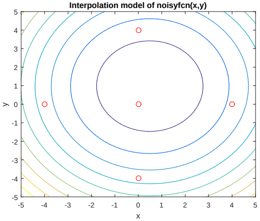






Derivative-Free Optimization - A Brief Intro





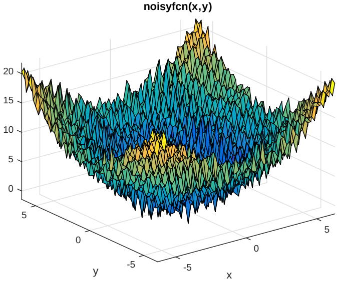
The gradient and Hessian of the model centered at x are *inexact* approximations of $\nabla f(x)$ and $\nabla^2 f(x)$ provided the model is

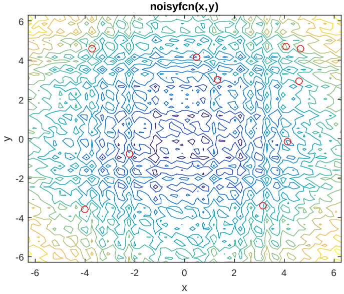
Definition (κ-fully linear.)

A function m is a κ-fully linear model of f on $\mathcal{B}(x, \Delta)$ provided, for $\kappa = (\kappa_{ef}, \kappa_{eg})$ and $\forall y \in \mathcal{B}(x, \Delta)$,

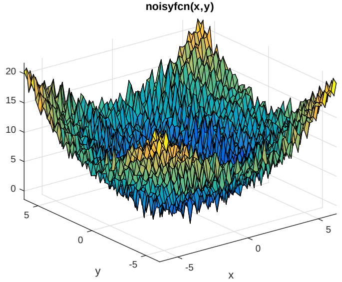
$$\begin{aligned} \|\nabla f(y) - \nabla m(y)\| &\leq \kappa_{eg}\Delta \text{ and} \\ |f(y) - m(y)| &\leq \kappa_{ef}\Delta^2 \end{aligned}$$

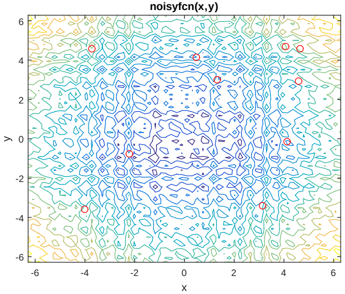
Random Models are Good, Too!

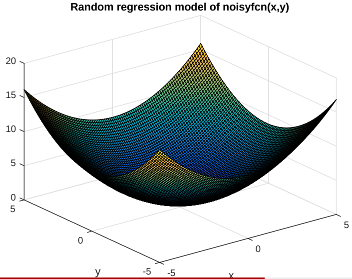


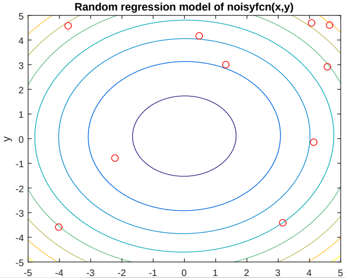


Random Models are Good, Too!









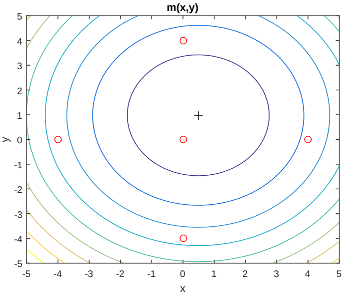
Model-Based DFO-TR Framework

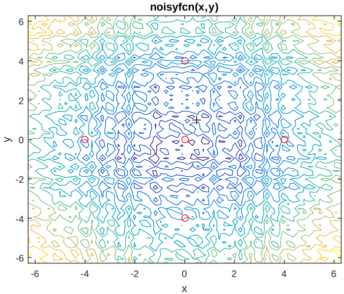
The heart of any DFO-TR method for unconstrained minimization:

Initialize x^0, $\Delta_0 > 0$, and some κ-fully linear gradient(Hessian) approximation $g_0(H_0)$.
While TRUE:

1. $s^* \leftarrow \arg \min_{s \in \mathcal{B}(0,\Delta_k)} m_k(s)$ where $m_k(s) = f(x^k) + g_k^T s + \frac{1}{2} s^T H_k s$
2. If $\dfrac{f(x^k) - f(x^k + s^*)}{m_k(x^k) - m_k(x^k + s^*)} > \eta > 0$, declare a *successful iteration*.
3. If successful, $x^{k+1} \leftarrow x^k + s^*$, $\Delta_{k+1} \geq \Delta_k$. Compute new κ-fully linear approximations g_{k+1}, H_{k+1} at x^k.
4. If unsuccessful, $x^{k+1} \leftarrow x^k$. At least one of $\Delta_{k+1} < \Delta_k$, compute new κ- fully linear approximations g_{k+1}, H_{k+1} at x^k.
5. $k \leftarrow k + 1$.

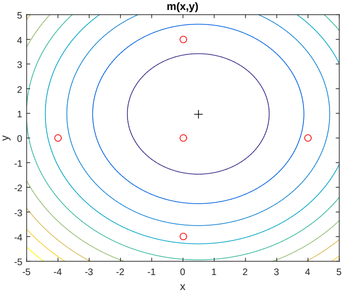
In Pictures - A Single Iteration and TR Subproblem

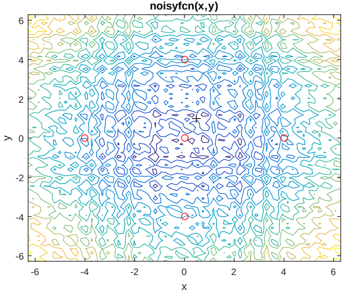




$$\text{Success ratio: } \frac{f(x^k) - f(x^k + s^*)}{m_k(x^k) - m_k(x^k + s^*)} > \eta > 0$$

In Pictures - A Single Iteration and TR Subproblem





$$\text{Success ratio: } \frac{f(x^k) - f(x^k + s^*)}{m_k(x^k) - m_k(x^k + s^*)} > \eta > 0$$

What if at each k we only have have an estimate f_k of $f(x^k)$ and f_k^+ of $f(x^k + x^+)$, generated by our estimator $f(x^k, \omega_k)$?

A Stochastic DFO-TR Framework

What if at each k we only have have an estimate f_k of $f(x^k)$ and f_k^+ of $f(x^k + x^+)$, generated by our estimator $f(x^k, \omega_k)$?

A Stochastic DFO-TR Framework

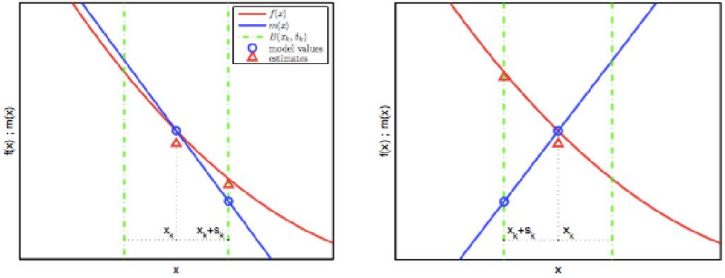
What if at each k we only have have an estimate f_k of $f(x^k)$ and f_k^+ of $f(x^k + x^+)$, generated by our estimator $f(x^k, \omega_k)$?

Initialize x^0, $\Delta_0 > 0$, and some random gradient(Hessian) approximation $g_0(H_0)$ at x^0.
While TRUE:

1. $s^* \leftarrow \arg \min_{s \in \mathcal{B}(0,\Delta_k)} m_k(s)$ where $m_k(s) = f(x^k) + g_k^T s + \frac{1}{2} s^T H_k s$
2. If $\dfrac{f_k - f_k^+}{m_k(x^k) - m_k(x^k + x^+)} > \eta > 0$, declare a *successful iteration*.
3. If successful, $x^{k+1} \leftarrow x^k + x^+$, $\Delta_{k+1} \geq \Delta_k$. Compute new random approximations g_{k+1}, H_{k+1} at x^{k+1}.
4. If unsuccessful, $x^{k+1} \leftarrow x^k$. ~~At least one of~~ $\Delta_{k+1} < \Delta_k$, compute new random approximations g_{k+1}, H_{k+1} at x^{k+1}.
5. $k \leftarrow k + 1$.

What Could Go Wrong - A Peek At Analysis

Recall, success determined by $\dfrac{f_k - f_k^+}{m_k(x^k) - m_k(x^k + x^+)} > \eta > 0$



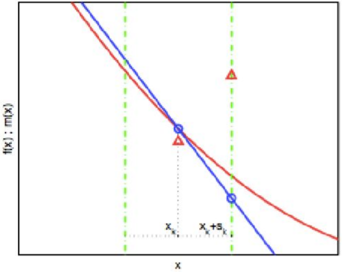
(a) Good model; good estimates. True successful steps.

(b) Bad model; good estimates. Unsuccessful steps.

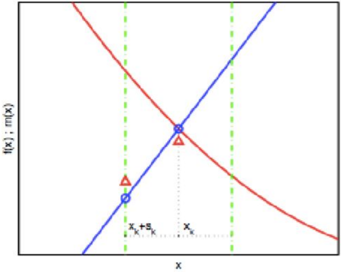
Images on this slide and next from Ruobing Chen's PhD thesis.

What Could Go Wrong - A Peek At Analysis

Recall, success determined by $\dfrac{f_k - f_k^+}{m_k(x^k) - m_k(x^k + x^+)} > \eta > 0$



(c) Good model; bad estimates.
Unsuccessful steps.



(d) Bad model; bad estimates.
False successful steps: f can increase!

Images on this slide and next from Ruobing Chen's PhD thesis.

Some Definitions

We just saw that we need good models and good estimates for success.
Moreover, we need to be very wary of FALSE successes!
Good models:

Some Definitions

We just saw that we need good models and good estimates for success. Moreover, we need to be very wary of FALSE successes!
Good models:

Definition

A sequence of random models $\{M_k\}$ is said to be α-probabilistically κ-fully linear with respect to the corresponding sequence $\{B(X_k, \Delta_k)\}$ if the events

$$I_k = \{M_k \text{ is a } \kappa\text{-fully linear model of } f \text{ on } B(X_k, \Delta_k)\}$$

satisfy the condition

$$P(I_k | \mathcal{F}_{k-1}^M) \geq \alpha,$$

where $\mathcal{F}_{k-1}^M$ is the σ-algebra generated by $M_0, \cdots, M_{k-1}$.

Some Definitions

We just saw that we need good models and good estimates for success. Good estimates:

Definition

The estimates f_k^0 and f_k^+ are said to be ϵ_F-accurate estimates of $f(x_k)$ and $f(x_k + x_k^+)$, respectively, for a given δ_k provided

$$|f_k^0 - f(x_k)| \leq \epsilon_F \delta_k^2 \text{ and } |f_k^+ - f(x_k + x_k^+)| \leq \epsilon_F \delta_k^2.$$

Some Definitions

We just saw that we need good models and good estimates for success. Good estimates:

Definition

The estimates f_k^0 and f_k^+ are said to be ϵ_F-accurate estimates of $f(x_k)$ and $f(x_k + x_k^+)$, respectively, for a given δ_k provided

$$|f_k^0 - f(x_k)| \leq \epsilon_F \delta_k^2 \text{ and } |f_k^+ - f(x_k + x_k^+)| \leq \epsilon_F \delta_k^2.$$

Definition

A sequence of random estimates $\{F_k^0, F_k^+\}$ is said to be β-probabilistically ϵ_F-accurate with respect to the corresponding sequence $\{X_k, \Delta_k, X_k^+\}$ if the events

$$J_k = \{F_k^0, F_k^+ \text{ are } \epsilon_F\text{-accurate estimates of } f(x_k) \text{ and } f(x_k + x_k^+), \text{ respectively, } \text{ for } \Delta_k\}$$

satisfy the condition

$$P(J_k | \mathcal{F}_{k-1/2}^{M \cdot F}) \geq \beta,$$

where ϵ_F is a fixed constant and $\mathcal{F}_{k-1/2}^{M \cdot F}$ is the σ-algebra generated by $M_0, \cdots, M_k$ and $F_0, \cdots, F_{k-1}$.

The Key Point

1. Model accuracy and estimate accuracy are both pegged to Δ_k.
2. Probabilities α, β are constants - noise can be "occasionally dominating"

The Key Point

1. Model accuracy and estimate accuracy are both pegged to Δ_k.
2. Probabilities α, β are constants - noise can be "occasionally dominating"
3. Analysis follows DFO-TR framework, additionally define r.v. $\Phi_k = \nu f(X_k) + (1-\nu)\Delta_k^2$, $\nu \in (0,1)$
4. Prove $\mathbb{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M \cdot F}] \leq -C\Delta_k^2 < 0$ (see the 1D pictures from before)

The Key Point

1. Model accuracy and estimate accuracy are both pegged to Δ_k.
2. Probabilities α, β are constants - noise can be "occasionally dominating"
3. Analysis follows DFO-TR framework, additionally define r.v. $\Phi_k = \nu f(X_k) + (1-\nu)\Delta_k^2$, $\nu \in (0,1)$
4. Prove $\mathbb{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M\cdot F}] \leq -C\Delta_k^2 < 0$ (see the 1D pictures from before)
5. So, Φ_k is a supermartingale. A bit more math, conclude from this that $\Delta_k \to 0$. By enforcing $\Delta_k < \eta_2 \|g^k\|$ on successful iterations,

The Key Point

1. Model accuracy and estimate accuracy are both pegged to Δ_k.
2. Probabilities α, β are constants - noise can be "occasionally dominating"
3. Analysis follows DFO-TR framework, additionally define r.v. $\Phi_k = \nu f(X_k) + (1-\nu)\Delta_k^2$, $\nu \in (0,1)$
4. Prove $\mathbb{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{M \cdot F}] \leq -C\Delta_k^2 < 0$ (see the 1D pictures from before)
5. So, Φ_k is a supermartingale. A bit more math, conclude from this that $\Delta_k \to 0$. By enforcing $\Delta_k < \eta_2 \|g^k\|$ on successful iterations,

Theorem (Rough Statement - Chen, M., Scheinberg 2015)

There exist $\alpha, \beta \in (0,1)$ and $\epsilon_F > 0$ dependent on f and algorithmic parameters so that if $\{M_k\}$ is α-probabilistically κ-fully linear and $\{F_k^0, F_k^+\}$ is β-probabilistically ϵ_F-accurate, then almost surely

$$\|\nabla f(X^k)\| \to 0.$$

A simple experiment - Function computation failures

Consider minimizing

$$f(x) = \sum_{i=1}^{n} (x_i - 1)^2.$$

but whenever for a given i, $|x_i - 1| < \epsilon$, we replace $(x_i - 1)^2$ with

$$f_i(x) = \begin{cases} (x_i - 1)^2 & \text{w.p. } 1 - \sigma \\ 10000 & \text{w.p. } \sigma \end{cases}$$

Use DFO-TR method with quadratic interpolation models.

Interpolation models built on random points within the current TR.

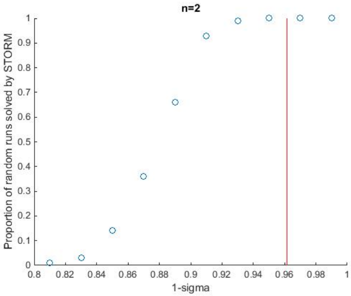
Initial point: $x^0 = 0$.

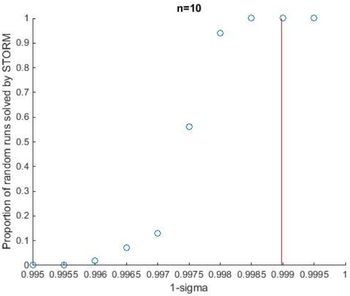
Good models : $\alpha \geq ((1-\sigma)^n)^{\frac{(n+1)(n+2)}{2}}$.

Good estimates : $\beta \geq ((1-\sigma)^n)^2$.

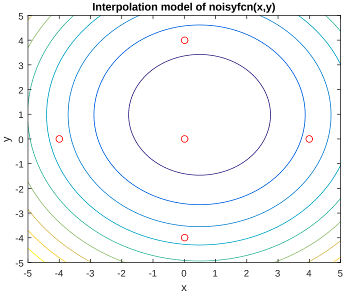
A simple experiment - Function computation failures

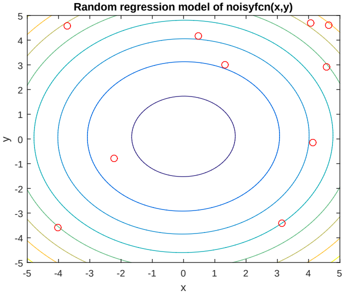
Comparing theory to practice:
Our theory predicts in red the least allowable value of $1 - \sigma$ for which our algorithm will guarantee convergence.





How do I construct my models?





- Larson, Billups (2013) - Poised sets for regression
- Shashaani, Hashemi, Pasupathy (2015) - ASTRO-DF - Poised sets for interpolation, adaptive sampling

- Scheinberg, M. (to appear) Uniform random sampling

α-probabilistically fully linear model sequences

Rough statement of a theorem:

Theorem

Let $\alpha \in (0,1)$. Let $\tilde{f}(x)$ be an unbiased estimator of $f(x)$ with standard deviation σ. Suppose you have p pairs $\{(x^i, \tilde{f}(x^i))\}_{i=1}^p$ where the x^i are drawn from a uniform distribution on $\mathcal{B}(0,\Delta)$ and $p \geq \max\{\kappa'/\Delta^4, 16(n+2)^2 \max\{2n, \ln(1/(1-\alpha))\}\}$. Let $\hat{w}$ denote the solution to

$$\min_w \sum_{i=1}^{p} (\tilde{f}(x^i) - \langle w, x^i \rangle)^2.$$

Then, with probability at least α,

$$\sup_{x \in \mathcal{B}(0,\Delta)} |f(x) - \langle \hat{w}, x \rangle| \leq \kappa \Delta^2,$$

where κ, κ' depend only on Lipschitz constants, n, σ, α, and numerical constants.

Conclusions and Future Work

Conclusions:

1. We proposed a method STORM for unconstrained stochastic optimization and proved a first-order stationarity result
2. Noise can occasionally be arbitrarily bad ("occasionally dominating")

Future/Ongoing Work:

1. Applying this work to various learning contexts (see: Katya's presentation)
2. Theoretical convergence rates? Rates for random model methods explored by Cartis and Scheinberg.