BAYESIAN OPTIMIZATION 101

Black box optimization – the sophisticated one

Diego Klabjan

Northwestern | McCORMICK SCHOOL OF ENGINEERING

Acknowledgment

- Images contributed by
  - Ryan P. Adams, School of Engineering and Applied Sciences, Harvard University

Key Concepts

- Function evaluated at certain points
- Non-evaluated points
  - Probabilistically estimate values
  - At any other point model the value as a distribution
    - Normal distribution
- Selecting the next point to evaluate
  - At each point we can calculate the expected value
    - Select a point with the lowest expected value

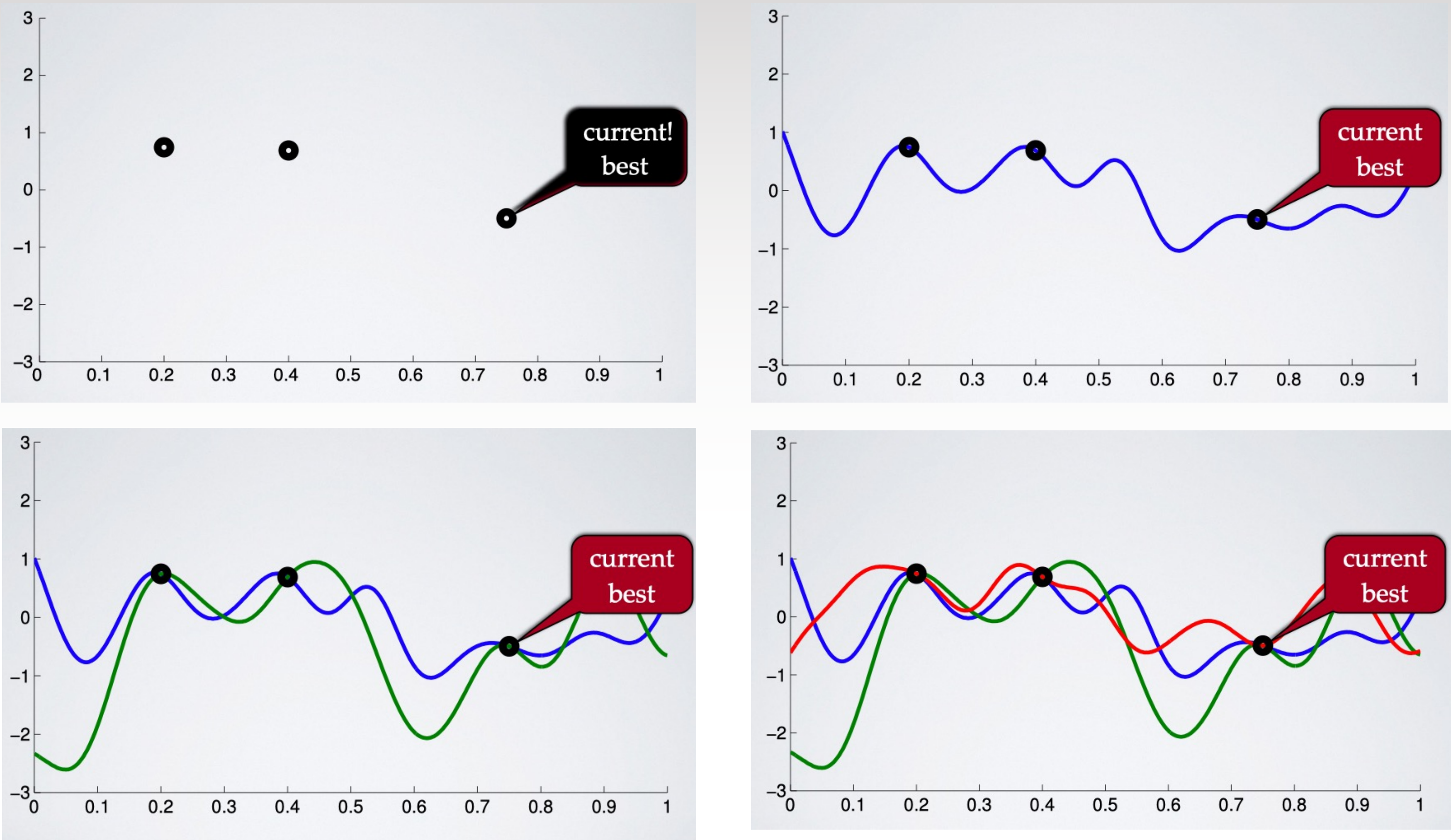
- However
  - Such a point might have high variance
    - Uncertainty
  - Lower uncertainty of points with high variance
- Trade-off
  - Low expected value
  - High variance

Gaussian Process

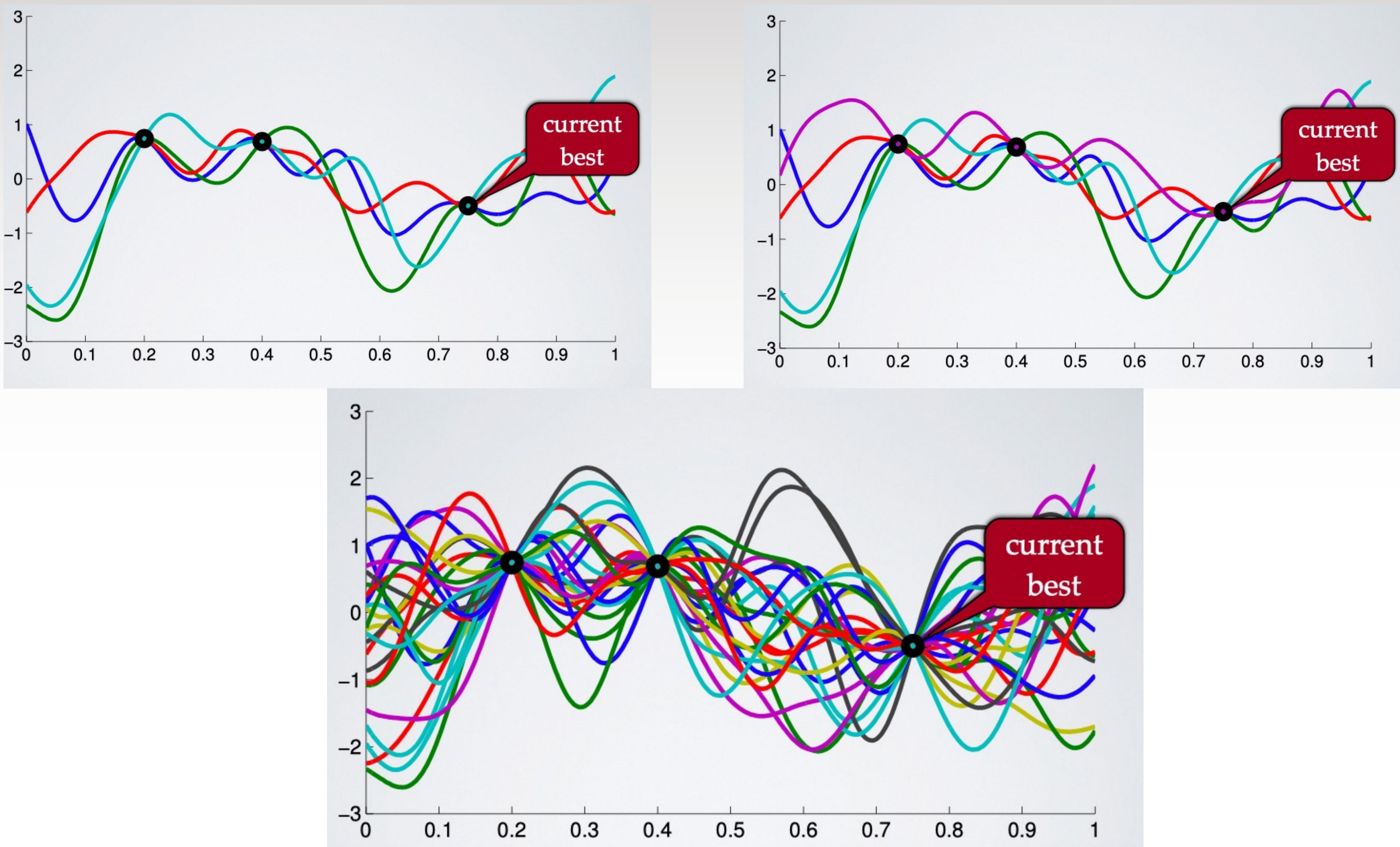
- Infinite and possibly uncountable many points

  $\{X_t : t \in T\}$

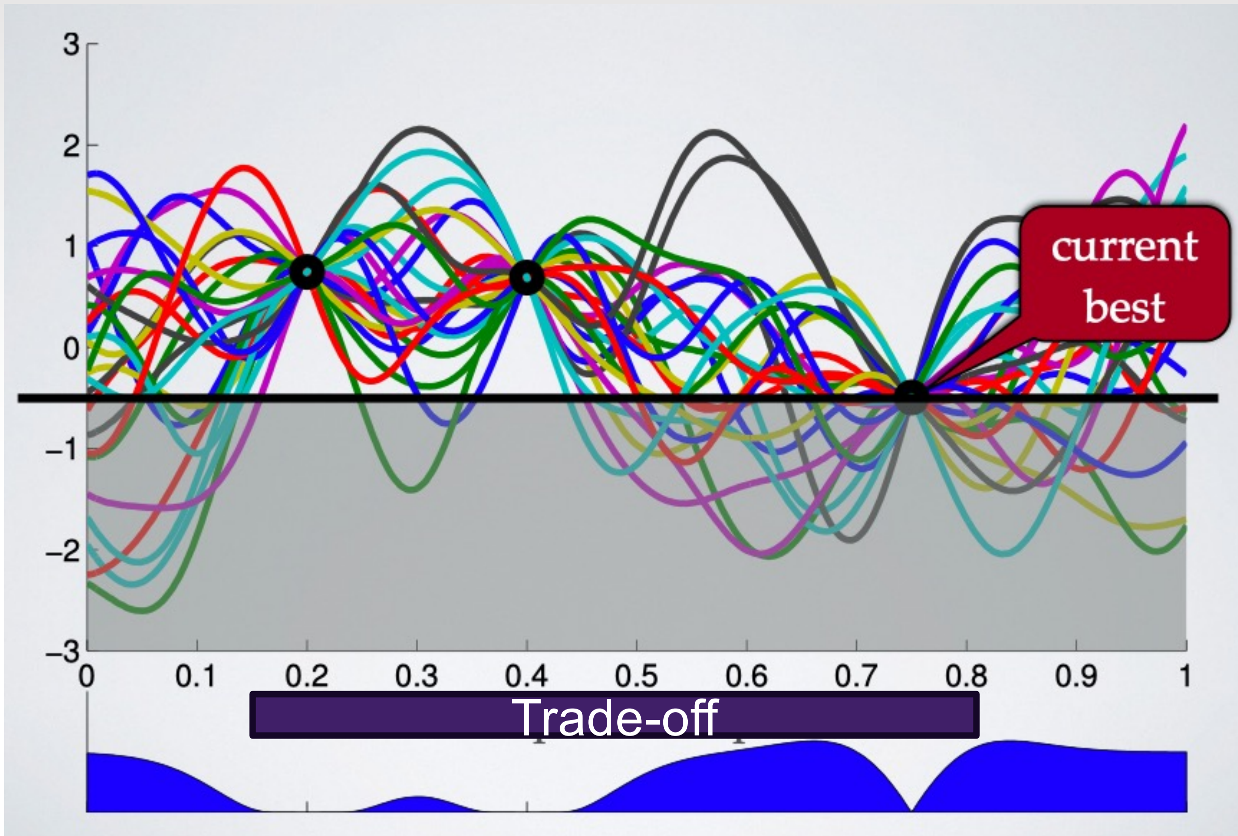
- Any finite subset of points distributed
  - Multi-variate Gaussian
- Equivalent to
  - Every linear combination is distributed based on single variate Gaussian


current!
best
current
best
current
best
current
best
3
2
1
0
−1
−2
−3
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

5



3
2
1
0
−1
−2
−3
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
current best
current best
current best


current best
Trade-off
3
2
1
0
−1
−2
−3
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Formalism

- Marginal likelihood

$$\ln p(\boldsymbol{y} \mid \boldsymbol{X}, \theta) = -\frac{N}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{K}_\theta| - \frac{1}{2} \boldsymbol{y}^\top \boldsymbol{K}_\theta^{-1} \boldsymbol{y}$$

- Test distribution on $\{x_m\}_1^M$
  - In BO used with $M = 1$

$$\boldsymbol{y}^{\text{test}} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma})$$

$$\boldsymbol{m} = \boldsymbol{k}_\theta^\top \boldsymbol{K}_\theta^{-1} \boldsymbol{y} \qquad \boldsymbol{\Sigma} = \boldsymbol{\kappa}_\theta - \boldsymbol{k}_\theta^\top \boldsymbol{K}_\theta^{-1} \boldsymbol{k}_\theta$$

Formalism

- Quantities
  - Covariance matrices

$$[\boldsymbol{K}_\theta]_{n,n'} = C(\boldsymbol{x}_n, \boldsymbol{x}_{n'}\,;\,\theta)$$

$$[\boldsymbol{k}_\theta]_{n,m} = C(\boldsymbol{x}_n, \boldsymbol{x}_m\,;\,\theta) \qquad [\boldsymbol{\kappa}_\theta]_{m,m'} = C(\boldsymbol{x}_m, \boldsymbol{x}_{m'}\,;\,\theta)$$

  - Parameter included if parametric covariance used
    - Notion of kernel
      - Kernels are usually parametric

Acquisition function

- Selection of next point to evaluate
- $\max_x m(x) - c\,\Sigma(x)$
- Max of mean – c standard deviation
  - Upper confidence bound
- Other acquisition functions possible
  - Probability of improvement
  - Expected improvement

Algorithm

- $S =$ set of a small number of random points
- Loop
  - Compute the covariance matrix $K = [u_{ij}] = \left[e^{-||x_i - x_j||}\right]_{x_i \in S, x_j \in S}$
  - Define vector $y = \left(f(x_i)\right)_i$
  - Define vector $k(x) = \left[e^{-||x - x_j||}\right]_{x_j \in S}$
  - Calculate $m(x) = k(x)K^{-1}y$
  - Calculate $\Sigma(x) = k(x) - k(x)K^{-1}k(x)$
  - Solve the acquisition problem
    - Let $\bar{x}$ be a solution
  - Evaluate $f(\bar{x})$
  - $S = S \cup \{\bar{x}\}$