Minimum Regret Search for Single- and Multi-Task Optimization

Jan Hendrik Metzen

University Bremen, Robotics Research Group
Robert Bosch GmbH, Corporate Research CR/AEY2

ICML, June 19, 2016

Introduction

- Motivation: optimization of expensive target functions
- Examples:
  - Automated machine learning (computational cost)
  - Process optimization (economical cost)
  - Robot learning (supervision cost)
  - Animal testing (moral cost)

Introduction

- Motivation: optimization of expensive target functions
- Examples:
  - Automated machine learning (computational cost)
  - Process optimization (economical cost)
  - Robot learning (supervision cost)
  - Animal testing (moral cost)
- Objective: Find close-to-optimal value of target function with a very small number of function evaluations ("queries")

Introduction

- Motivation: optimization of expensive target functions
- Examples:
  - Automated machine learning (computational cost)
  - Process optimization (economical cost)
  - Robot learning (supervision cost)
  - Animal testing (moral cost)
- Objective: Find close-to-optimal value of target function with a very small number of function evaluations ("queries")
- Approach: Invest significant amounts of computation time to determine "optimal" sequence of query points

Bayesian Optimization

Bayesian optimization[1] in a nutshell:

- black-box optimization problems: $\mathbf{x} = \arg\max_{\mathbf{x} \in \mathcal{X}} f(x)$ of some function $f : \mathcal{X} \to \mathbb{R}$ on some bounded set $\mathcal{X} \subset \mathbb{R}^D$.
- **probabilistic model** $p(f)$ for $f(\mathbf{x})$, typically a Gaussian process (GP)

[1] B. Shahriari, K. Swersky, Z. Wang, R. Adams, and N. de Freitas. Taking the Human Out of the Loop: A Review of Bayesian Optimization

Bayesian optimization[1] in a nutshell:

- black-box optimization problems: $\mathbf{x} = \arg\max_{\mathbf{x}\in\mathcal{X}} f(x)$ of some function $f : \mathcal{X} \to \mathbb{R}$ on some bounded set $\mathcal{X} \subset \mathbb{R}^D$.
- **probabilistic model** $p(f)$ for $f(\mathbf{x})$, typically a Gaussian process (GP)
- For $n = 1 \ldots N$:
  - determine GP posterior $p(f|\mathcal{D}_n)$ for $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$
  - decide on a query point based on **acquisition function** a: $\mathbf{x}_{n+1} = \arg\max_{\mathbf{x}\in\mathcal{X}} a_{p(f|\mathcal{D}_n)}(x)$
  - observe (potentially noisy) $y_{n+1} = f(\mathbf{x}_{n+1}) + \epsilon$

[1] B. Shahriari, K. Swersky, Z. Wang, R. Adams, and N. de Freitas. Taking the Human Out of the Loop: A Review of Bayesian Optimization

Bayesian optimization[1] in a nutshell:

- black-box optimization problems: $\mathbf{x} = \arg\max_{\mathbf{x} \in \mathcal{X}} f(x)$ of some function $f : \mathcal{X} \rightarrow \mathbb{R}$ on some bounded set $\mathcal{X} \subset \mathbb{R}^D$.
- **probabilistic model** $p(f)$ for $f(\mathbf{x})$, typically a Gaussian process (GP)
- For $n = 1 \dots N$:
  - determine GP posterior $p(f|\mathcal{D}_n)$ for $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$
  - decide on a query point based on **acquisition function** a: $\mathbf{x}_{n+1} = \arg\max_{\mathbf{x} \in \mathcal{X}} a_{p(f|\mathcal{D}_n)}(x)$
  - observe (potentially noisy) $y_{n+1} = f(\mathbf{x}_{n+1}) + \epsilon$
- recommend $\tilde{\mathbf{x}}_N$ as optimum after N queries (optimum of GP or best query point)
- objective: minimize **simple regret** $R_f(\tilde{\mathbf{x}}_N) = f(\mathbf{x}^\star) - f(\tilde{\mathbf{x}}_N) = \max_{\mathbf{x}} f(\mathbf{x}) - f(\tilde{\mathbf{x}}_N)$

[1] B. Shahriari, K. Swersky, Z. Wang, R. Adams, and N. de Freitas. Taking the Human Out of the Loop: A Review of Bayesian Optimization

Entropy Search

- Let $p^\star(x|\mathcal{D}_n)$ denote the posterior distribution (after observing $\mathcal{D}_n$) of the unknown optimizer $\mathbf{x}^\star = \arg\max_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$.
- and $H(\mathbf{x}^\star|\mathcal{D}_n)$ denote the differential entropy of $p^\star(x|\mathcal{D}_n)$
- Entropy Search (ES)[2] is an information theoretic acquisition fct.:

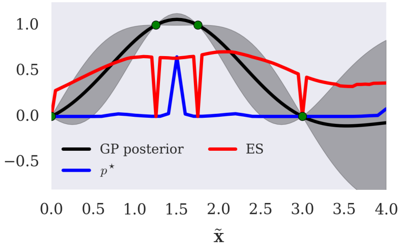
$$a_{ES}(\mathbf{x}, \mathcal{D}_n) = \underbrace{H(\mathbf{x}^\star|\mathcal{D}_n)}_{\text{current entropy}} - \underbrace{\mathbb{E}_{y|\mathbf{x},\mathcal{D}_n}[H(\mathbf{x}^\star|\mathcal{D}_n \cup \{(\mathbf{x}, y)\})]}_{\text{expected posterior entropy for query at } \mathbf{x}}$$

[2] P. Hennig and C. J. Schuler. Entropy Search for Information-Efficient Global Optimization. *JMLR*, 13:1809–1837, 2012

Entropy Search

- Let $p^\star(x|\mathcal{D}_n)$ denote the posterior distribution (after observing $\mathcal{D}_n$) of the unknown optimizer $\mathbf{x}^\star = \arg\max_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$.
- and $H(\mathbf{x}^\star|\mathcal{D}_n)$ denote the differential entropy of $p^\star(x|\mathcal{D}_n)$
- Entropy Search (ES)[2] is an information theoretic acquisition fct.:

$$a_{ES}(\mathbf{x}, \mathcal{D}_n) = \underbrace{H(\mathbf{x}^\star|\mathcal{D}_n)}_{\text{current entropy}} - \underbrace{\mathbb{E}_{y|\mathbf{x},\mathcal{D}_n}[H(\mathbf{x}^\star|\mathcal{D}_n \cup \{(\mathbf{x}, y)\})]}_{\text{expected posterior entropy for query at } \mathbf{x}}$$

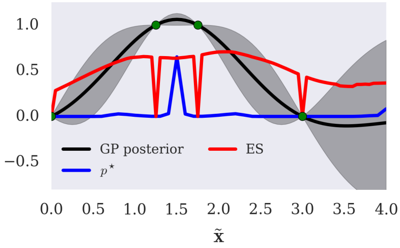


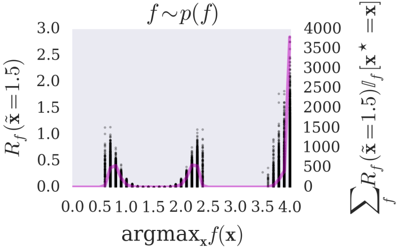
[2] P. Hennig and C. J. Schuler. Entropy Search for Information-Efficient Global Optimization. *JMLR*, 13:1809–1837, 2012

Entropy Search

- Let $p^\star(x|\mathcal{D}_n)$ denote the posterior distribution (after observing $\mathcal{D}_n$) of the unknown optimizer $\mathbf{x}^\star = \arg\max_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$.
- and $H(\mathbf{x}^\star|\mathcal{D}_n)$ denote the differential entropy of $p^\star(x|\mathcal{D}_n)$
- Entropy Search (ES)[2] is an information theoretic acquisition fct.:

$$a_{ES}(\mathbf{x},\mathcal{D}_n) = \underbrace{H(\mathbf{x}^\star|\mathcal{D}_n)}_{\text{current entropy}} - \underbrace{\mathbb{E}_{y|\mathbf{x},\mathcal{D}_n}[H(\mathbf{x}^\star|\mathcal{D}_n \cup \{(\mathbf{x},y)\})]}_{\text{expected posterior entropy for query at } \mathbf{x}}$$



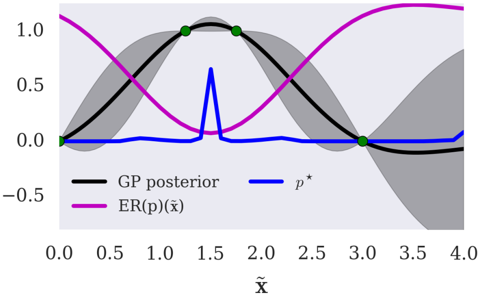


[2] P. Hennig and C. J. Schuler. Entropy Search for Information-Efficient Global Optimization. *JMLR*, 13:1809–1837, 2012

Minimum Regret Search (MRS)

Acquisition function based on minimizing the expected simple regret

- Expected simple regret:
  $\mathrm{ER}(p)(\mathbf{x}) = \mathbb{E}_{p(f)}[R_f(\mathbf{x})] = \mathbb{E}_{p(f)}[\max_{\mathbf{x}} f(\mathbf{x}) - f(\mathbf{x})]$
- For fixed GP $p(f)$, $\tilde{\mathbf{x}} = \arg\min_{\mathbf{x}} \mathrm{ER}(p)(\mathbf{x})$ corresponds to the maximizer of the GP mean



Minimum Regret Search (MRS)

Acquisition function based on minimizing the expected simple regret

- Expected simple regret:
  $\mathrm{ER}(p)(\mathbf{x}) = \mathbb{E}_{p(f)}[R_f(\mathbf{x})] = \mathbb{E}_{p(f)}[\max_{\mathbf{x}} f(\mathbf{x}) - f(\mathbf{x})]$
- MRS aims at selecting query points s.t. ER is minimized also with respect to resulting $p(f)$
- Myopic choice: $\mathrm{MRS}^{\mathrm{point}}$ selects next query point s.t. minimum ER is reduced the most (in expectation)

$$a_{\mathrm{MRS}^{\mathrm{point}}}(\mathbf{x}^q) = \underbrace{\min_{\tilde{\mathbf{x}}} \mathrm{ER}(p_n)(\tilde{\mathbf{x}})}_{\text{current minimum ER}} - \underbrace{\mathbb{E}_{y|p_n(f),\mathbf{x}^q}[\min_{\tilde{\mathbf{x}}} \mathrm{ER}(p_n^{[\mathbf{x}^q,y]})(\tilde{\mathbf{x}})]}_{\text{expected posterior minimum ER for query at } \mathbf{x}^q}$$

Minimum Regret Search (MRS)

Acquisition function based on minimizing the expected simple regret

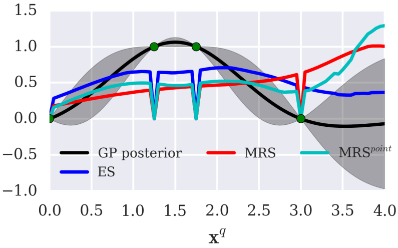
- Expected simple regret:
  $\mathrm{ER}(p)(\mathbf{x}) = \mathbb{E}_{p(f)}[R_f(\mathbf{x})] = \mathbb{E}_{p(f)}[\max_{\mathbf{x}} f(\mathbf{x}) - f(\mathbf{x})]$
- MRS aims at selecting query points s.t. ER is minimized also with respect to resulting $p(f)$
- Myopic choice: $\mathrm{MRS}^{\mathrm{point}}$ selects next query point s.t. minimum ER is reduced the most (in expectation)

$$a_{\mathrm{MRS}^{\mathrm{point}}}(\mathbf{x}^q) = \underbrace{\min_{\tilde{\mathbf{x}}} \mathrm{ER}(p_n)(\tilde{\mathbf{x}})}_{\text{current minimum ER}} - \underbrace{\mathbb{E}_{y|p_n(f),\mathbf{x}^q}[\min_{\tilde{\mathbf{x}}} \mathrm{ER}(p_n^{[\mathbf{x}^q,y]})(\tilde{\mathbf{x}})]}_{\text{expected posterior minimum ER for query at } \mathbf{x}^q}$$

- Minimizer $\arg\min_{\tilde{\mathbf{x}}} \mathrm{ER}(p_n)(\tilde{\mathbf{x}})$ can be seen as point estimate for $\tilde{\mathbf{x}}_N$
- MRS additionally also accounts for uncertainty regarding $\tilde{\mathbf{x}}_N$:

$$a_{\mathrm{MRS}}(\mathbf{x}^q) = \mathbb{E}_{\tilde{\mathbf{x}}\sim p^\star_{\mathcal{D}_n}}[\mathrm{ER}(p_n)(\tilde{\mathbf{x}})] - \mathbb{E}_{y|p_n(f),\mathbf{x}^q}[\mathbb{E}_{\tilde{\mathbf{x}}\sim p^\star_{\mathcal{D}_n\cup\{(\mathbf{x}^q,y)\}}}[\mathrm{ER}(p_n^{[\mathbf{x}^q,y]})(\tilde{\mathbf{x}})]$$

Illustration of Acquisition Functions



- ES tends to query close to areas where $p^\star$ is large
- MRS$^{\text{point}}$ tends to query in areas which are risky for current recommendation (large simple regret possible)
- MRS is more smooth than MRS$^{\text{point}}$ since it accounts for uncertainty in recommendation

Experimental Setup

Synthetic Single-Task Benchmark[3]:

- Target functions sampled from a generative model on $\mathcal{X} = [0, 1]^2$
- In practice:
    - sample 250 pairs $(\mathbf{x}, f(\mathbf{x}))$ from function $f \sim p(f)$
    - fit GP to these pairs
    - use resulting posterior mean as target function

[3] P. Hennig and C. J. Schuler. Entropy Search for Information-Efficient Global Optimization. *JMLR*, 13:1809–1837, 2012

Experimental Setup

Synthetic Single-Task Benchmark[3]:

- Target functions sampled from a generative model on $\mathcal{X} = [0, 1]^2$
- In practice:
  - sample 250 pairs $(\mathbf{x}, f(\mathbf{x}))$ from function $f \sim p(f)$
  - fit GP to these pairs
  - use resulting posterior mean as target function
- Gaussian noise with standard deviation $\sigma = 10^{-3}$ is added to each observation
- Probabilistic surrogate model used in BO:
  GP with isotropic RBF kernel of length scale $l = 0.1$ and unit signal variance

[3] P. Hennig and C. J. Schuler. Entropy Search for Information-Efficient Global Optimization. *JMLR*, 13:1809–1837, 2012

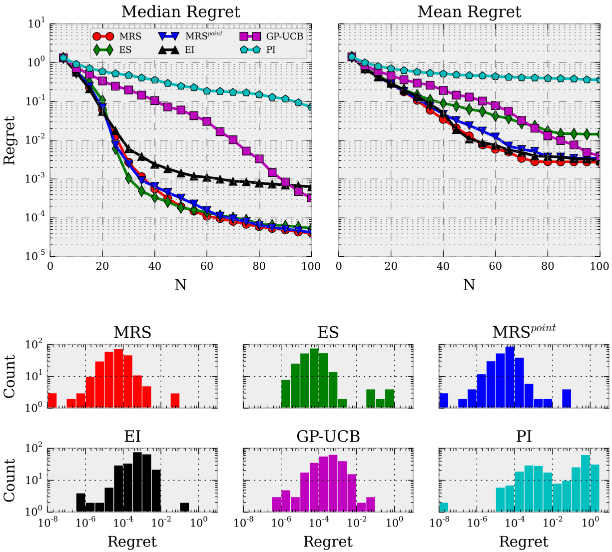
Experimental Setup

Synthetic Single-Task Benchmark[3]:

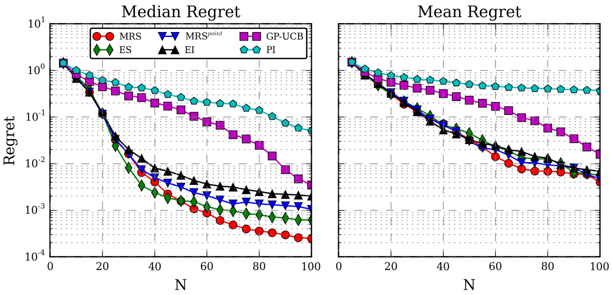
- Target functions sampled from a generative model on $\mathcal{X} = [0,1]^2$
- In practice:
  - sample 250 pairs $(\mathbf{x}, f(\mathbf{x}))$ from function $f \sim p(f)$
  - fit GP to these pairs
  - use resulting posterior mean as target function
- Gaussian noise with standard deviation $\sigma = 10^{-3}$ is added to each observation
- Probabilistic surrogate model used in BO:
  GP with isotropic RBF kernel of length scale $l = 0.1$ and unit signal variance
- Generative model $p(f)$
  - without model mismatch: GP with isotropic RBF kernel ($l = 0.1$ and unit signal variance)
  - with model mismatch: GP with isotropic rational quadratic kernel ($l = 0.1$, $\alpha = 1.0$ and unit signal variance)

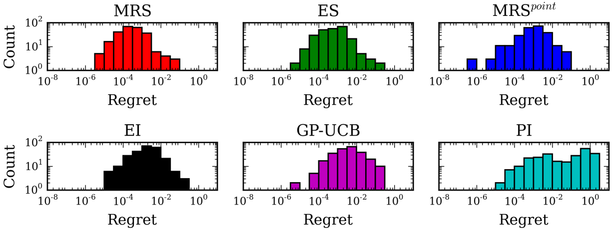
[3] P. Hennig and C. J. Schuler. Entropy Search for Information-Efficient Global Optimization. *JMLR*, 13:1809–1837, 2012

Results: No model-mismatch



Results: Model-mismatch





Summary and Outlook

- Contribution: Novel acquisition function which explicitly aims at minimizing the expected simple regret

Summary and Outlook

- Contribution: Novel acquisition function which explicitly aims at minimizing the expected simple regret
- Additional content of the paper:
  - multi-task minimum regret search
  - results on simulated robotic task

Summary and Outlook

- Contribution: Novel acquisition function which explicitly aims at minimizing the expected simple regret
- Additional content of the paper:
  - multi-task minimum regret search
  - results on simulated robotic task
- Future work:
  - treatment of GP hyperparameters
  - more efficient approximation techniques for MRS

Summary and Outlook

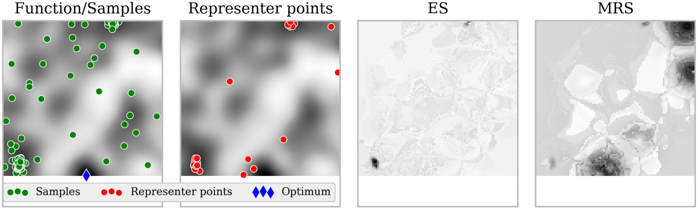
- Contribution: Novel acquisition function which explicitly aims at minimizing the expected simple regret
- Additional content of the paper:
  - multi-task minimum regret search
  - results on simulated robotic task
- Future work:
  - treatment of GP hyperparameters
  - more efficient approximation techniques for MRS
- Source code available at:
  `https://github.com/jmetzen/bayesian_optimization`

Summary and Outlook

- Contribution: Novel acquisition function which explicitly aims at minimizing the expected simple regret
- Additional content of the paper:
  - multi-task minimum regret search
  - results on simulated robotic task
- Future work:
  - treatment of GP hyperparameters
  - more efficient approximation techniques for MRS
- Source code available at:
  `https://github.com/jmetzen/bayesian_optimization`

Thank you for your attention and see you at the poster!
Do you have questions, comments, or ideas?

Results: MRS versus ES



Acquisition functions on a target function at $N = 100$ and 25 representer points; darker areas correspond to larger values. ES focuses on sampling in areas with high density of $p^\star$ (many representer points), while MRS focuses on unexplored areas that are populated by representer points (non-zero $p^\star$).