SC4/SM8 Advanced Topics in Statistical Machine Learning

Chapter 10: Bayesian Optimisation

Dino Sejdinovic
Department of Statistics
Oxford

Slides and other materials available at:
http://www.stats.ox.ac.uk/~sejdinov/atsml19/

Optimizing “black-box” functions

Machine learning models often have a number of hyperparameters which need to be tuned:

- **kernel methods**: bandwidth in a Gaussian kernel, degree of a Matérn kernel, regularization parameters
- **neural networks**: number of layers, regularization parameters, layer size, batch size, learning rate
- **Latent Dirichlet Allocation**: Dirichlet parameters, number of topics, vocabulary size

Define an objective function: a measure of generalization performance for a given set of hyperparameters obtained e.g. using cross-validation.

- Grid search, random search, trial-and-error...

Optimizing “black-box” functions

We are interested in optimizing a 'well behaved' function $f : \mathcal{X} \to \mathbb{R}$ over some bounded domain $\mathcal{X} \subset \mathbb{R}^d$, i.e. in solving

$$x_\star = \mathrm{argmin}_{x \in \mathcal{X}} f(x).$$

However, f is not known explicitly, i.e. it is a **black-box** function and we can only ever obtain noisy (and potentially expensive as they may correspond to training of a large machine learning model or even running a complex physical experiment) evaluations of f.

Probabilistic model for the objective f

- Assuming that f is well behaved, we build a surrogate probabilistic model for it (Gaussian Process).
- Compute the posterior predictive distribution of f
- Optimize a cheap proxy / acquisiton function instead of f which takes into account predicted values of f at new points as well as the uncertainty in those predictions: this model is typically much cheaper to evaluate than the actual objective f.
- Evaluate the objective f at the optimum of the proxy.

The proxy / acquisiton function should balance exploration against exploitation.

Surrogate GP model

Assume that the noise ϵ_i in the evaluations of the black-box function is i.i.d. $\mathcal{N}\left(0, \delta^2\right)$:

$$\mathbf{f} \sim \mathcal{N}(0, \mathbf{K})$$
$$\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \delta^2 I).$$

Gives a closed form expression for the **posterior predictive mean** $\mu\left(x\right)$ and the **posterior predictive marginal standard deviation** $\sigma\left(x\right) = \sqrt{\kappa\left(x, x\right)}$ at any new location x, i.e.

$$f\left(x\right)|\mathcal{D} \sim \mathcal{N}\left(\mu\left(x\right), \kappa\left(x, x\right)\right),$$

where

$$\begin{aligned}
\mu\left(x\right) &= \mathbf{k}_{x\mathbf{x}}(\mathbf{K} + \delta^2 I)^{-1}\mathbf{y}, \\
\kappa\left(x, x\right) &= k(x, x) - \mathbf{k}_{x\mathbf{x}}(\mathbf{K} + \delta^2 I)^{-1}\mathbf{k}_{\mathbf{x}x}
\end{aligned}$$

- Exploitation: seeking locations with low posterior mean $\mu\left(x\right)$,
- Exploration: seeking locations with high posterior variance $\kappa\left(x, x\right)$.

Acquisition functions

- **GP-LCB**. "optimism in the phase of uncertainty"; minimize the lower $(1-\alpha)$-credible bound of the posterior of the unknown function values $f(x)$, i.e.

$$\alpha_{LCB}\left(x\right)=\mu\left(x\right)-z_{1-\alpha}\sigma\left(x\right),$$

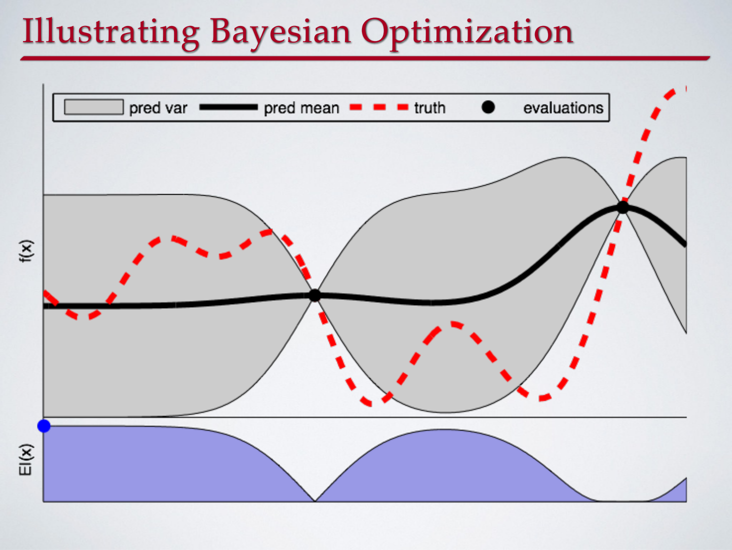
  where $z_{1-\alpha}=\Phi^{-1}\left(1-\alpha\right)$ is the desired quantile of the standard normal distibution.
- **PI** (probability of improvement). $\tilde{x}$: the optimal location so far, $\tilde{y}$: the observed minimum. Let $u\left(x\right)=\mathbb{1}\left\{f\left(x\right)<\tilde{y}\right\}$,

$$\alpha_{PI}\left(x\right)=\mathbb{E}\left[u(x)|\mathcal{D}\right]=\Phi\left(\gamma(x)\right),\quad \gamma(x)=\frac{\tilde{y}-\mu\left(x\right)}{\sigma\left(x\right)}$$

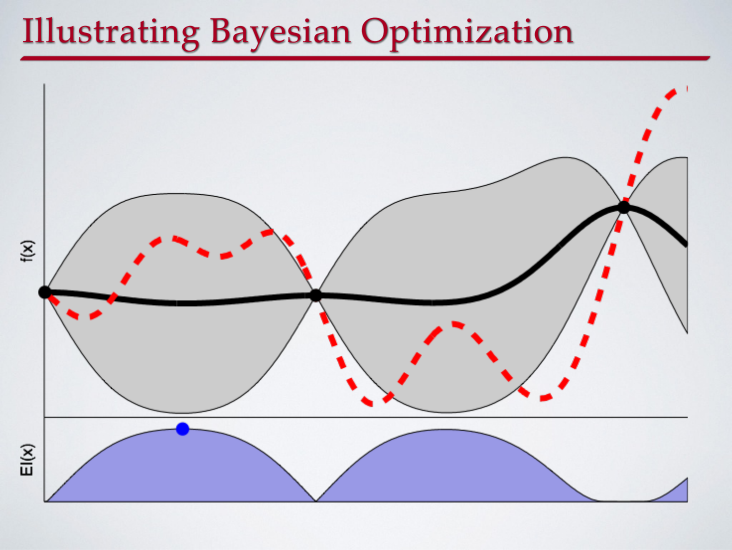
- **EI** (expected improvement). Let $u\left(x\right)=\max\left(0,\tilde{y}-f\left(x\right)\right)$

$$\alpha_{EI}\left(x\right)=\mathbb{E}\left[u(x)|\mathcal{D}\right]=\sigma\left(x\right)\left(\gamma\left(x\right)\Phi\left(\gamma\left(x\right)\right)+\phi\left(\gamma(x)\right)\right).$$

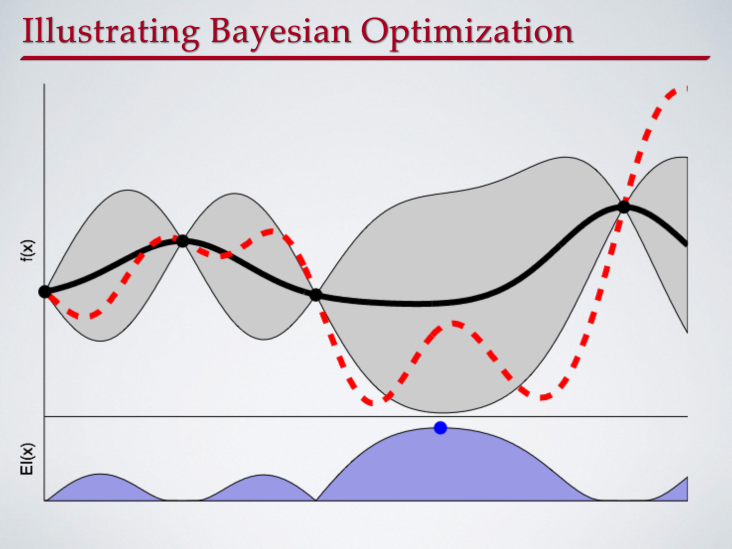
Treating $\tilde{y}$ as the actual value $f(\tilde{x})$ of the objective?



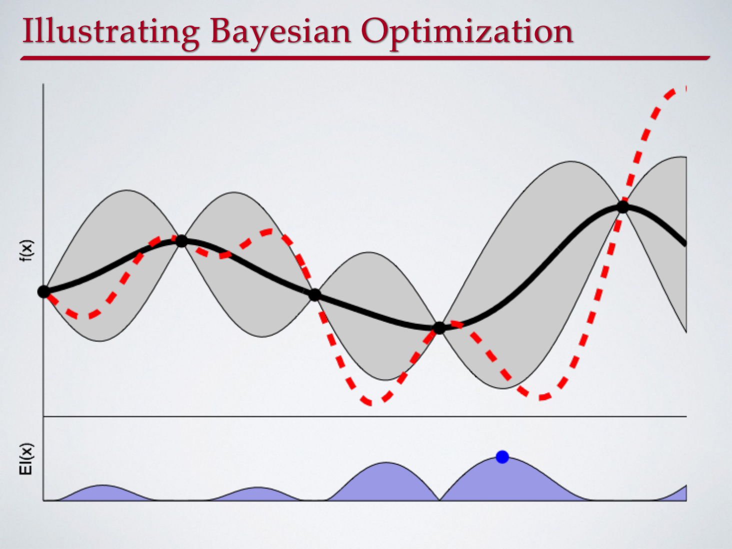
slides from **A Tutorial on Bayesian Optimization for Machine Learning** by Ryan Adams



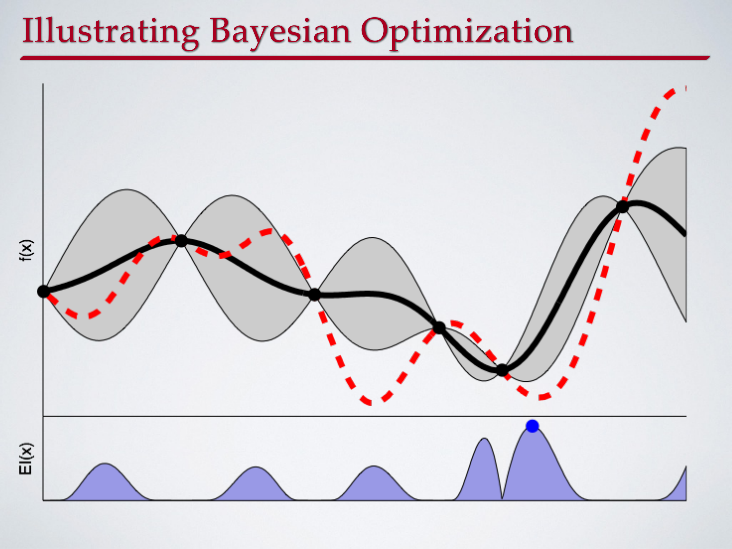
slides from **A Tutorial on Bayesian Optimization for Machine Learning** by Ryan Adams



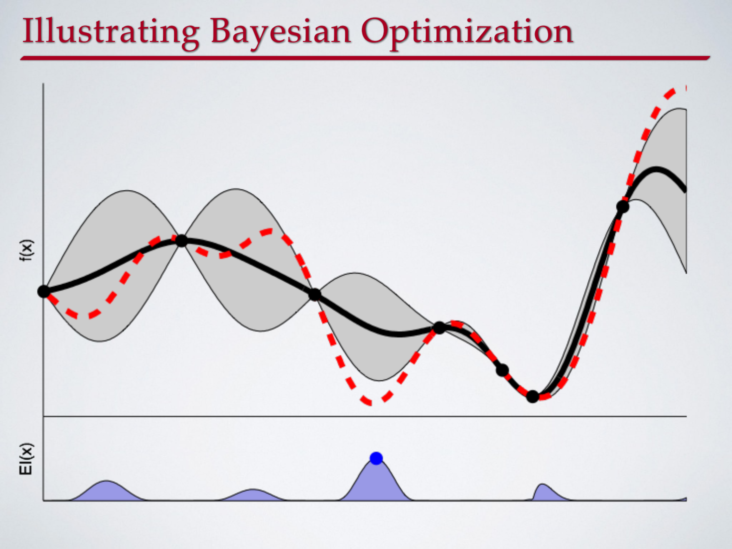
slides from **A Tutorial on Bayesian Optimization for Machine Learning** by Ryan Adams



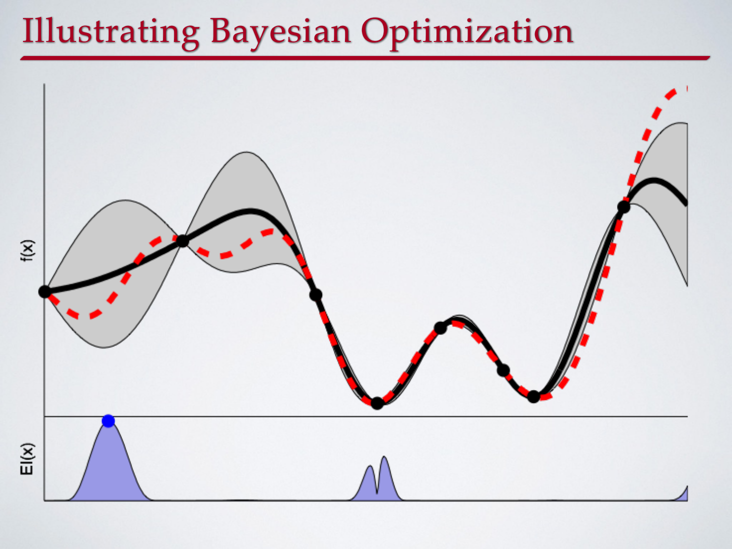
slides from **A Tutorial on Bayesian Optimization for Machine Learning** by Ryan Adams



slides from **A Tutorial on Bayesian Optimization for Machine Learning** by Ryan Adams



slides from **A Tutorial on Bayesian Optimization for Machine Learning** by Ryan Adams



slides from **A Tutorial on Bayesian Optimization for Machine Learning** by Ryan Adams

We considered a selection of topics in statistical machine learning, but there is much more!

- **Topics we did not cover**: multitask learning, online learning, reinforcement learning, deep learning, message passing algorithms, generative adversarial networks, ensemble methods, boosting, causality, interpretability, robustness, fairness, differential privacy...
- **Further resources**:
  - Bishop, Pattern Recognition and Machine Learning, Springer.
  - Murphy, Machine Learning: A Probabilistic Perspective, MIT Press.
  - Shalev-Shwartz and Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press.
  - Schölkopf and Smola, Learning with Kernels, MIT Press.
  - Rasmussen and Williams, Gaussian Processes for Machine Learning, MIT Press.
  - Goodfellow, Bengio and Courville, Deep Learning, MIT Press.
  - Machine Learning Summer Schools, videolectures.net.
  - Conferences: NeurIPS, ICML, AISTATS, UAI.