Efficient Optimization with Gray-box simulations

January 5, 2016

Explain:

1. What is the goal? Optimizatin, twin model advantage of cheap gradient, reuse samples
2. Why not sample g
3. What am I going to show? Why x^3.

1 Posterior with fixed covariance parameters

The unknown true model

$$f(x)$$

The gradient of the twin model is

$$\nabla g(x) = \nabla f(x) + \epsilon(x)\,, \tag{1}$$

where $\epsilon(x)$ is an unkown realization of Gaussian process $\mathcal{N}(0, cov_2(\cdot,\cdot))$.

$$E: \quad cov_2(\epsilon(x_1), \epsilon(x_2)) = \xi_2^2 \boldsymbol{I} \exp\left\{-\frac{(x_1 - x_2)^2}{2\sigma_2^2}\right\} \tag{2}$$

The true model is modeled as a realization of

$$f \sim \mathcal{N}\left(\bar{f}_D, cov_1(\cdot,\cdot)\right)\,, \tag{3}$$

where

$$A_{11}: \quad cov_1(f(x_1), f(x_2)) = \xi_1^2 \exp\left\{-\frac{(x_1 - x_2)^2}{2\sigma_1^2}\right\} \tag{4}$$

and $\bar{f}_D$ is the sample mean. Assume

$$\begin{aligned} cov(\nabla f, \epsilon) &= 0 \\ cov(f, \epsilon) &= 0 \end{aligned} \tag{5}$$

Thus

$$A_{13}: \quad cov(f, \nabla g) = cov(f, \nabla f) \tag{6}$$

Therefore

$$A_{12} = A_{13}: \quad cov(f(x_1), \nabla f(x_2)) = \frac{\xi_1^2}{\sigma_1^2}(x_1 - x_2) \exp\left\{-\frac{(x_1 - x_2)^2}{2\sigma_1^2}\right\} \tag{7}$$

$$A_{22}: \quad cov(\nabla f(x_1), \nabla f(x_2)) = \frac{\xi_1^2}{\sigma_1^2} \exp\left\{-\frac{(x_1 - x_2)^2}{2\sigma_1^2}\right\} \left(\boldsymbol{I} - \frac{1}{\sigma_1^2}(x_1 - x_2)(x_1 - x_2)^T\right) \tag{8}$$

$$A_{23} = A_{22}: \quad cov(\nabla f, \nabla g) = cov(\nabla f, \nabla f) \tag{9}$$

$$A_{33}: \quad cov(\nabla g, \nabla g) = cov(\nabla f, \nabla f) + cov(\epsilon, \epsilon) \tag{10}$$

We have

$$\begin{pmatrix} f \\ \nabla f \\ \nabla g \end{pmatrix} = \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\bar{f}_D} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \Sigma = \begin{pmatrix} A_{11} & A_{12} & A_{12} \\ A_{12}^T & A_{22} & A_{22} \\ A_{12}^T & A_{22} & A_{22} + E \end{pmatrix}\right) \tag{11}$$

The Gaussian processes are discretized at X. Denote X_D as the sampled points where f and ∇g are available. Define $d = (f(X_D), \nabla g(X_D))$. Permute the rows and columns such that Σ_D is the matrix corresponding to $f(X_D)$ and $\nabla g(X_D)$. The permuted covariance matrix is

$$\begin{pmatrix} \Sigma_{\backslash D} & \Sigma_{\backslash DD} \\ \Sigma_{\backslash DD}^T & \Sigma_D \end{pmatrix} \tag{12}$$

Then

$$\begin{aligned} \mu_{\backslash D|D} &= \mu_{\backslash D} + \Sigma_{\backslash DD}\Sigma_D^{-1}(d - \mu_D) \\ \Sigma_{\backslash D|D} &= \Sigma_{\backslash D} - \Sigma_{\backslash DD}\Sigma_D^{-1}\Sigma_{\backslash DD}^T \end{aligned} \tag{13}$$

Compare the posterior f between using approximate gradient and without using approximate gradient. Compare the posterior ∇f between using approximate gradient and without using approximate gradient.

2 Estimating covariance parameters

3 Acquisition Function

3.1 Probability of Improvement

3.2 Expected Improvement

3.3 Lower Confidence Bound

Minimize $f(x)$. Consider minimizing the acquisition function

$$\min \mu(x) - \kappa\sigma(x) \tag{14}$$

The minimizer dictates the next point to sample. Suppose the trust-region is $[-2, 2]$.

4 Trust-region Optimization

When the dimension is high, expressing the GP, computing its posterior, and maximizing the acquisition function in this high-dimensional ball is expensive. I am thinking of construct the mesh for GP according to a cone around the approximate gradient. Switch between gradient-free, gradient-driven based on quality of twin model's gradient.

Trying to show: modeling posterior of true model and use Bayesian optimization provide a natural switch. Next step: prove if gradient approximation is good enough, then the probability that Bayesian optimization approach dictates the next sample point the same as gradient-driven methods (which one? BFGS, grad-descent, ...?) goes to 1 (clearly the convergence rate depends on the trust region size and Hessian). In the other extreme, the probability of dictate a point the same as with just the sampling the function value using Bayesian optimization goes to 1.

5 Finding the next sample point

Denote x as the point to sample, D as the sampled data points. We have

$$\begin{pmatrix} f_x \\ \boldsymbol{f_D} \\ \boldsymbol{\nabla g_D} \end{pmatrix} \sim \mathcal{N}\left\{ \begin{pmatrix} \bar{f}_D \\ \boldsymbol{\bar{f}_D} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} cov_1(f(x), f(x)) & cov_1(f(x), f(x_D)) & cov_1(f(x), \nabla f(x_D)) \\ cov_1(f(x), f(x_D))^T & A_{11} & A_{12} \\ cov_1(f(x), \nabla f(x_D))^T & A_{12}^T & A_{22} + E \end{pmatrix} \right\} \tag{15}$$

Therefore,

$$\mu_x|_D = \bar{f}_D + \left(cov_1(f_x, f_D), cov_1(f_x, \nabla f_D)\right) \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} + E \end{pmatrix}^{-1} \begin{pmatrix} d_{f_D} - \bar{f}_D \\ d_{\nabla g_D} \end{pmatrix} \tag{16}$$

$$\sigma_x^2|_D = cov_1(f_x, f_x) - \left(cov_1(f_x, f_D), cov_1(f_x, \nabla f_D)\right) \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} + E \end{pmatrix}^{-1} \begin{pmatrix} cov_1(f_x, f_D) \\ cov_1(f_x, \nabla f_D) \end{pmatrix} \tag{17}$$

Therefore,

$$\frac{\partial \mu_x|_D}{\partial x} = \left(cov_1(\nabla f_x, f_D), cov_1(\nabla f_x, \nabla f_D)\right) \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} + E \end{pmatrix}^{-1} \begin{pmatrix} d_{f_D} - \bar{f}_D \\ d_{\nabla g_D} \end{pmatrix} \tag{18}$$

$$\frac{\partial \sigma_x^2|_D}{\partial x} = cov_1(\nabla f_x, f_x) - 2\left(cov_1(\nabla f_x, f_D), cov_1(\nabla f_x, \nabla f_D)\right) \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} + E \end{pmatrix}^{-1} \begin{pmatrix} cov_1(f_x, f_D) \\ cov_1(f_x, \nabla f_D) \end{pmatrix} \tag{19}$$

$$\frac{\partial \sigma_x|_D}{\partial x} = \frac{1}{2\sqrt{\sigma_x^2|_D + \delta^2}} \frac{\partial \sigma_x^2|_D}{\partial x} \tag{20}$$

5.1 Expected Improvement

If we use the expected improvement

$$I(x) = \mathbb{E}\left\{\max\left(0, f(x) - f(x^+)\right)\right\} \tag{21}$$

Then

$$\mathbb{E}I = \sigma(x)\left\{\frac{\mu(x) - f(x^+)}{\sigma(x)}\right\}\Phi\left(\frac{\mu(x) - f(x^+)}{\sigma(x)}\right) + \phi\left(\frac{\mu(x) - f(x^+)}{\sigma(x)}\right) \tag{22}$$

where ϕ is the PDF, Φ is the CDF of standard normal distribution.

5.2 Rosenbrock function

$$f(\mathbf{x}) = \sum_{i=0}^{N-2} 100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2$$

$$\frac{\partial f}{\partial x_j} = 200(x_j - x_{j-1}^2) - 400x_j(x_{j+1} - x_j^2) + 2(x_j - 1) \quad \text{for } j = 1, \cdots, N-2$$

$$\frac{\partial f}{\partial x_0} = -400x_0(x_1 - x_0^2) + 2(x_0 - 1)$$

$$\frac{\partial f}{\partial x_{N-1}} = 200(x_{N-1} - x_{N-2}^2)$$