Stochastic Gradient Descent for Gaussian Processes Done Right

Jihao Andreas Lin*, Shreyas Padhy*, Javier Antorán*, Austin Tripp, Alexander Terenin, Csaba Szepesvári, José Miguel Hernández-Lobato, David Janz

Three things you NEED to know to solve large quadratic problems with stochastic gradient descent!

1. Gaussian Process Regression

Data set: $X \in \mathbb{R}^{n \times d}, y \in \mathbb{R}^N$

Define $\alpha^\star(b) = (K + \lambda I)^{-1} b$ Kernel function: k, kernel matrix: $K \in \mathbb{R}^{n \times n}$

Posterior mean: $k(\,\cdot\,, X)\, \alpha^\star(y)$ Likelihood variance: $\lambda > 0$, prior sample: f_0

Posterior sample: $f_0(\,\cdot\,) + k(\,\cdot\,, X)\, \alpha^\star(y - (f_0(X) + \zeta))$ $\zeta \sim \mathcal{N}(0, \lambda I)$

2. Use ~~Primal~~ Dual Objective

$\alpha^\star = \arg\min_\alpha L(\alpha)$

primal $L(\alpha) = \frac{1}{2}\|b - K\alpha\|_2^2 + \frac{\lambda}{2}\|\alpha\|_K^2$

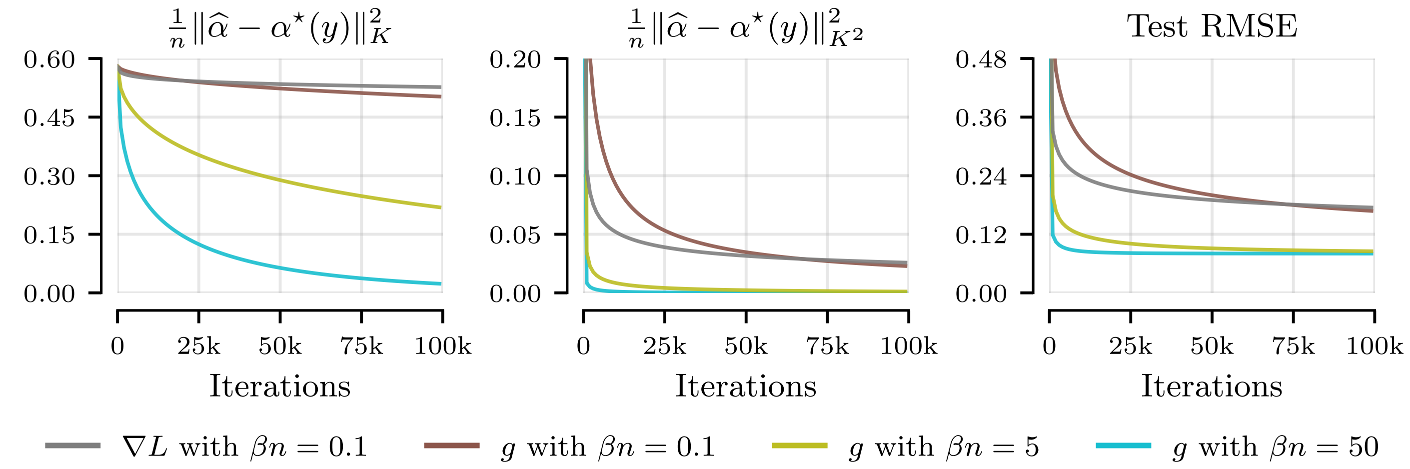
$\nabla L(\alpha) = K(\lambda\alpha - b + K\alpha)$

$\nabla^2 L = K(K + \lambda I)$

dual $L^*(\alpha) = \frac{1}{2}\|\alpha\|_{K+\lambda I}^2 - \alpha^\top b$

$g(\alpha) = \nabla L^*(\alpha) = \lambda\alpha - b + K\alpha$

$\nabla^2 L^* = K + \lambda I$



3. Use Random ~~Features~~ Coordinates

$g(\alpha) = \lambda\alpha - b + K\alpha$

random features $\mathbb{E}[m z_j z_j^\top] = K$

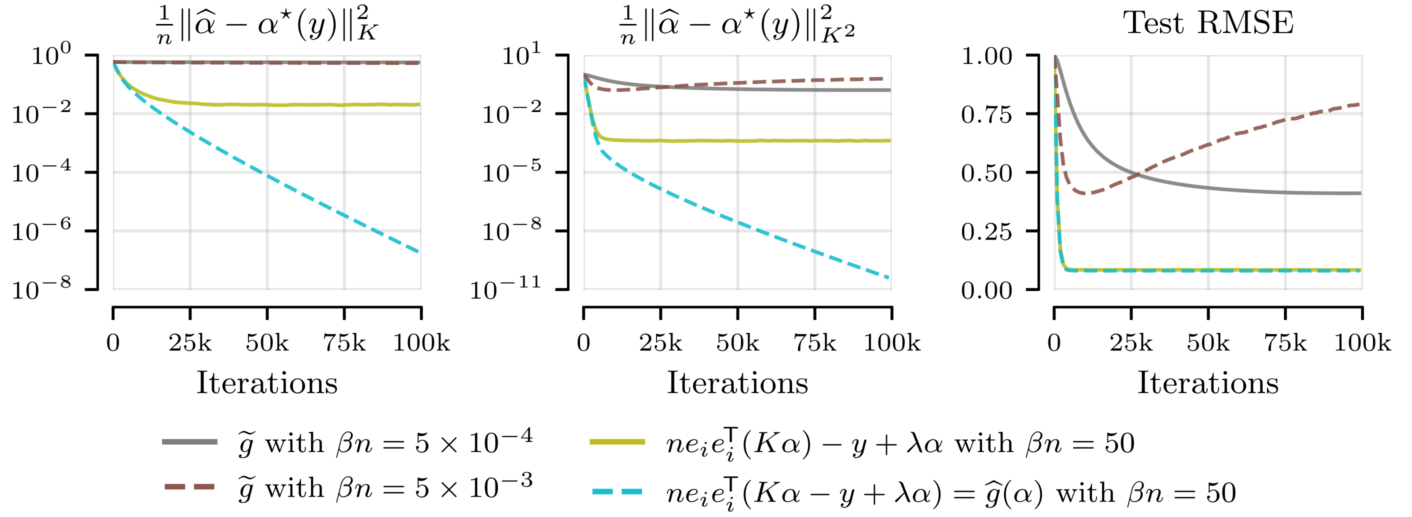
$\widetilde{g}(\alpha) = \lambda\alpha - b + m z_j z_j^\top \alpha$

additive noise

random coordinates $\mathbb{E}[n e_i e_i^\top] = I$

$\widehat{g}(\alpha) = n e_i e_i^\top(\lambda\alpha - b + K\alpha)$

multiplicative noise



4. Use Momentum and Geometric Averaging

Nesterov's momentum

$v_t = \rho v_{t-1} - \beta g(\alpha_{t-1} + \rho v_{t-1})$

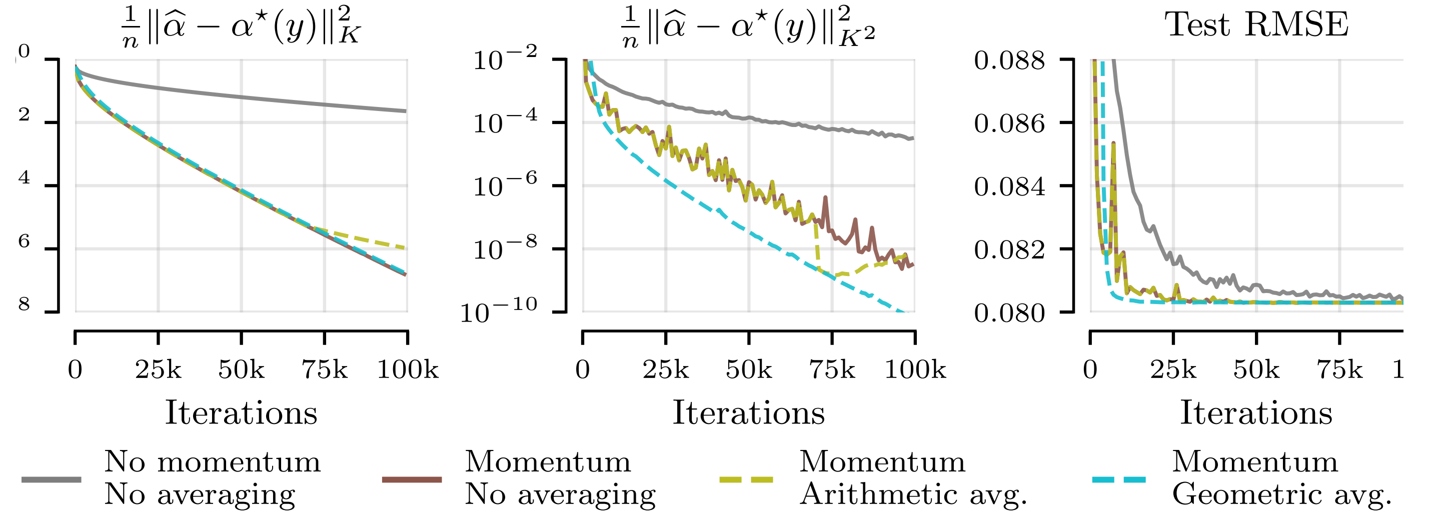
$\alpha_t = \alpha_{t-1} + v_t$

arithmetic averaging

$\bar{\alpha}_t = \frac{1}{t - t_0}\sum_{l=t_0}^{t} \alpha_l$

geometric averaging

$\bar{\alpha}_t = r\alpha_t + (1 - r)\bar{\alpha}_{t-1}$



5. Parallel Thompson Sampling

synthetic data, $d = 8$