ISBN: 978-3-95450-227-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2022-WEPOST030

MULTITASK OPTIMIZATION OF LASER-PLASMA ACCELERATORS USING SIMULATION CODES WITH DIFFERENT FIDELITIES

A. Ferran Pousa[1], S. Jalas[2], M. Kirchen[1], A. Martinez de la Ossa[1], M. Thévenet[1], S. Hudson[3], J. Larson[3], A. Huebl[4], J.-L. Vay[4], R. Lehe[4]
[1]Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, Hamburg, Germany
[2]Department of Physics Universität Hamburg, Luruper Chaussee 149, Hamburg, Germany
[3]Argonne National Laboratory, Lemont, IL, USA
[4]Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Abstract

When designing a laser-plasma acceleration experiment, one commonly explores the parameter space (plasma density, laser intensity, focal position, etc.) with simulations in order to find an optimal configuration that, for example, minimizes the energy spread or emittance of the accelerated beam. However, laser-plasma acceleration is typically modeled with full particle-in-cell (PIC) codes, which can be computationally expensive. Various reduced models can approximate beam behavior at a much lower computational cost. Although such models do not capture the full physics, they could still suggest promising sets of parameters to be simulated with a full PIC code and thereby speed up the overall design optimization.

In this work we automate such a workflow with a Bayesian multitask algorithm, where each task has a different fidelity. This algorithm learns from past simulation results from both full PIC codes and reduced PIC codes and dynamically chooses the next parameters to be simulated. We illustrate this workflow with a proof-of-concept optimization using the Wake-T and FBPIC codes. The libEnsemble library is used to orchestrate this workflow on a modern GPU-accelerated high-performance computing system.

INTRODUCTION

Laser-plasma accelerators (LPAs) are a promising acceleration technology that could have applications in high-energy physics, medicine, and materials science [1]. Many of these applications require finely tuning the different parameters of a given setup (e.g., plasma density, laser intensity, beam profile) in order to attain optimal performance (e.g., optimal beam quality). This *design optimization* is usually done by running a particle-in-cell (PIC) simulation for each prospective set of parameters in order to assess its corresponding performance (often quantified by a single objective function such as the final beam energy spread). Because these simulations are computationally expensive, however, one would like to make informed choices about which set of parameters to evaluate, so as to find the optimal configuration with as few simulations as possible.

One method to find these parameters is Bayesian optimization [2], whereby a Gaussian process model [3] of the objective function over the parameter space is progressively learned. At each iteration of this method, the model suggests the most promising set of parameters to be assessed by the simulation, and the simulation result is in turn used to update and refine the model. Importantly, the feasibility of Bayesian optimization for automated tuning of LPA setups has recently been demonstrated [4, 5].

A further step to reduce computational cost is to perform some of the simulations at a lower fidelity. Indeed, a number of *reduced* codes for laser-plasma acceleration have been developed that make different types of approximations. These approximations can result in dramatically faster simulations, at the cost of a potential loss of accuracy. Examples of these approximations include assuming cylindrical symmetry [6], averaging over the fast laser oscillations [7], or assuming the wakefield to be quasi-static [7, 8].

Here we show that incorporating lower-fidelity simulation output from reduced PIC codes into a Bayesian optimization method can reduce the overall computational cost of obtaining a high-fidelity solution by an order of magnitude. The combination of PIC codes with different fidelities into a single optimization is enabled by the multitask Bayesian optimization (MTBO) algorithm [9, 10], a special case of multifidelity optimization that operates with two levels of fidelity (i.e., two *tasks*): an inexpensive, low-fidelity model for broad parameter exploration and a computationally demanding, high-fidelity model for which only a reduced number of well-targeted simulations are performed. Incidentally, we note that other types of multifidelity algorithms have also been shown to perform well in a number of problems, including with multiobjective optimization [11].

MULTITASK BAYESIAN OPTIMIZATION

The MTBO algorithm [9, 10] builds a Gaussian process model [3] of the simulation output (a scalar that quantifies beam quality in our case) as a function of both the vector of design parameters x (e.g., plasma density, beam profile parameters) and the fidelity d. (We let $d = 1$ denote low-fidelity and $d = 2$ the high-fidelity, and we denote the respective simulation output at a given fidelity by $f_d(x)$.) Accordingly, the correlation kernel used inside the Gaussian process model depends on both the parameters x and the fidelity d and is assumed to be of the form $k((d, \mathbf{x}), (d', \mathbf{x}')) = B_{dd'}\kappa(x - x')$, where κ is typically a Mattérn kernel [3] and B is a 2×2 symmetric matrix. In practice, the coefficients of B (as well as the parameters of κ) are hyperparameters that are automatically determined by maximizing the marginal likelihood of the

ISBN: 978-3-95450-227-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2022-WEPOST030

previously observed data. In particular, B_{12} quantifies the apparent level of correlation between low- and high-fidelity results.

Given this model, the algorithm performs an iterative loop whereby, at each iteration, the following occurs:

- n_1 "promising" points in parameter space $\{x_i\}_{i=1,..,n_1}$ are chosen by maximizing an acquisition function (typically, the expected improvement [3]) based on the Gaussian process model for the *high-fidelity* output $f_2(x)$.
- These n_1 points $\{x_i\}_{i=1,..,n_1}$ are evaluated by using the *low-fidelity* simulations, and the Gaussian process model is updated based on these new simulation results.
- The updated Gaussian process model for f_2 is evaluated on the n_1 original points $\{x_i\}_{i=1,..,n_1}$, and the n_2 points (with $n_2 < n_1$) that give the highest values are selected.
- These n_2 points are evaluated by using the *high-fidelity* simulations, and the Gaussian process model is updated based on these new simulation results.

Thus, unlike other methods that *dynamically* choose the fidelity of each evaluation, the MTBO algorithm uses fixed-size batches for low-fidelity and high-fidelity evaluations.

PROOF-OF-PRINCIPLE STUDY

The effectiveness of the multitask approach is demonstrated here by means of a proof-of-principle optimization study combining the simulation codes FBPIC [12] and Wake-T [13]. While FBPIC provides a high-fidelity, fully electromagnetic PIC description of the LPA physics in quasi-3D geometry [6], Wake-T allows for inexpensive simulations by using a reduced quasi-static wakefield model with 2D cylindrical symmetry [14] and an envelope model [15] for the laser driver.

The setup to be optimized is an LPA stage acting as an energy booster for an externally injected beam. Given a fixed laser driver, the goal of the optimizer is to tune the current profile of the electron beam so that beamloading minimizes the energy spread while maintaining high charge [16–18]. More specifically, although the wakefield varies in time because of dephasing, depletion, and diffraction, an optimal current profile must be found that, on average, results in uniform acceleration along the beam. This issue is generally addressed with simulations or directly in experiments [4, 19].

In this study the current profile of the beam is chosen to be trapezoidal (known to be the optimal profile for an idealized plasma bubble [17]) and is defined by four optimization parameters: the current at the head, I_h; the current at the tail, I_t; the length of the beam, L_b; and its longitudinal position in the wake, parameterized as $\Delta z_\mathrm{l,h} = z_l - z_\mathrm{h}$, which corresponds to the distance between the position of the head of the beam, z_h, and the center of the laser driver, z_l. To achieve both low energy spread and high charge, we combine these two quantities into a single objective function to maximize: $f = k_Q E_\mathrm{MED}[\mathrm{GeV}]/k_\mathrm{MAD}$, where $k_Q = Q_\mathrm{tot}/Q_\mathrm{ref}$ is the ratio between the total beam charge Q_tot and a reference charge $Q_\mathrm{ref} = 10\,\mathrm{pC}$ and where $k_\mathrm{MAD} = \Delta E_\mathrm{MAD}/\Delta E_\mathrm{MAD,ref}$ is the ratio between the beam relative energy spread ΔE_MAD, measured as the median absolute deviation (MAD), and a reference $\Delta E_\mathrm{MAD,ref} = 0.01$. E_MED is the median energy of the beam. The use of MAD and MED variables provides a robust measure of the energy distribution of LPA beams, which typically feature long, low-charge energy tails [4, 19].

The laser driver considered for this study has an energy $E_L = 10\,\mathrm{J}$, a FWHM duration $\tau_\mathrm{FWHM} = 25\,\mathrm{fs}$, a spot size $w_0 = 40\,\mu\mathrm{m}$, a wavelength $\lambda_0 = 800\,\mathrm{nm}$, and a peak normalized vector potential $a_0 \simeq 2.6$. The plasma density profile is a simple 10 cm-long flat-top profile with an on-axis electron density $n_{e,0} = 2 \times 10^{17}\,\mathrm{cm}^{-3}$ and a parabolic profile in the radial direction for laser guiding $n_e(r) = n_{e,0} + r^2/(\pi r_e w_0^4)$. The externally injected electron beam has an initial energy $E_{b,0} = 200\,\mathrm{MeV}$ with an RMS energy spread of 0.1 %. Its transverse phase-space is elliptical, featuring normalized emittances of $\epsilon_{n,x} = 3\,\mu\mathrm{m}$ (horizontal) and $\epsilon_{n,y} = 0.5\,\mu\mathrm{m}$ (vertical). This difference between the horizontal and vertical emittances can typically be observed in beams from LPAs based on ionization injection as a result of the laser polarization [4]. The beam parameters exposed to the optimizer can vary in the following predefined ranges: $I_h \in [0.1, 10]\mathrm{kA}$, $I_t \in [0.1, 10]\mathrm{kA}$, $L_b \in [1, 20]\mu\mathrm{m}$, and $\Delta z_{l,h} \in [40, 60]\mu\mathrm{m}$.

The FBPIC simulations are performed by using the boosted frame technique [20, 21] with a Lorentz boost factor of 25. They have longitudinal and radial resolutions of $dz = \lambda_0/80$ and $dr = 0.6\,\mu\mathrm{m}$, respectively, and use three azimuthal modes in order to properly describe the ellipticity of the particle beam. Each simulation is performed on a single NVIDIA A100 GPU with a typical execution time of ~45 min. The Wake-T simulations feature longitudinal and transverse resolutions of $dz = \tau_\mathrm{FWHM}/20$ and $dr = 0.6\,\mu\mathrm{m}$, respectively. Each simulation is performed on a single core of an AMD EPYC 7643 CPU with a typical execution time of 4 min to 6 min. The entire multitask optimization is carried out on one compute node with 96 CPU cores and 4 GPUs. One GPU is reserved for the optimizer, which uses the Ax implementation of the MTBO algorithm [22], while the other three are allocated for FBPIC simulations. With this setup, the optimizer is able to perform batches of either $n_1 = 90$ concurrent low-fidelity simulations with Wake-T or $n_2 = 3$ concurrent high-fidelity simulations with FBPIC. The allocation of GPU and CPU resources to the different simulations, as well as the coordination and communication between the simulations and the optimizer is handled by the libEnsemble library [23].

The typical evolution of a multitask optimization run is visualized in Fig. 1(a). The large number of Wake-T simulations allows for broad parameter exploration, so that only the most promising configurations are evaluated with FBPIC. Although the outcomes of the Wake-T and FBPIC simulations do not fully agree, the clear correlation between them (see Fig. 1(b)) allows the optimizer to gain valuable information from the low-fidelity simulation output.

ISBN: 978-3-95450-227-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2022-WEPOST030

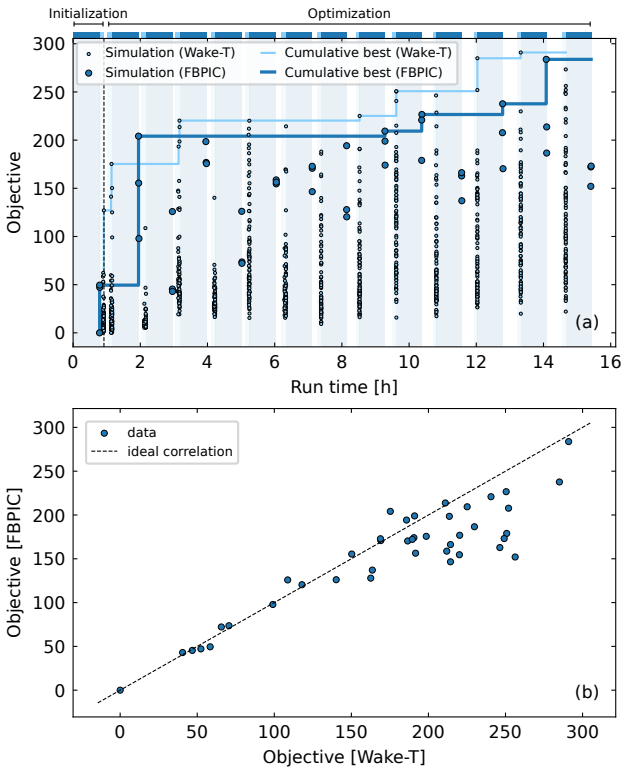

Figure 1: (a) Visualization of the Wake-T and FBPIC simulation batches carried out during a multitask optimization run, including the outcome of each simulation and the evolution of the cumulative best objective in both fidelities. (b) Correlation between the outcome of the FBPIC evaluations and their Wake-T counterparts for the same run as in (a).

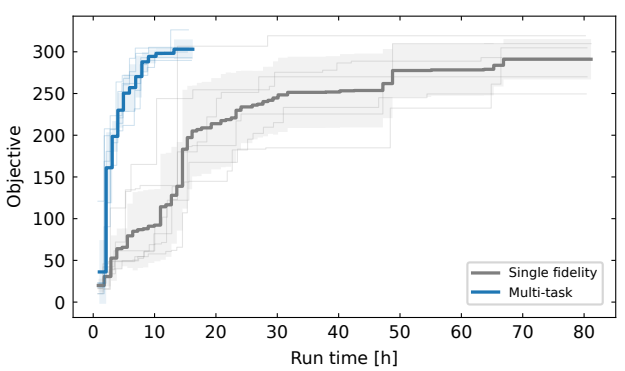

Figure 2: Average (thick line) and standard deviation (shaded area) of the evolution of the objective function $f = k_Q E_{MED}/k_{MAD}$ with the multitask and single-fidelity algorithms. Six runs (thin lines) were performed for each case. The multitask line includes the values of only the objective evaluated at high fidelity (FBPIC).

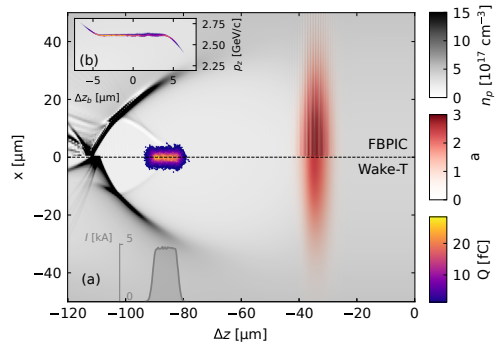

Figure 3: (a) Plasma wakefields 5 cm into the LPA as obtained from FBPIC (top) and Wake-T (bottom) for the optimal set of parameters. (b) Longitudinal phase space at the end of the FBPIC simulation.

For comparison, a series of Bayesian optimization runs with a standard single-fidelity algorithm are also performed. These runs consist exclusively of FBPIC simulations and are performed on the same hardware (i.e., batches of 3 simulations running on NVIDIA A100 GPUs). For both the multi-fidelity and single-fidelity cases, the optimization starts with a first batch of simulations using randomly chosen points within the parameter space [24]. Thus, each optimization run evolves differently. To account for this effect, we performed 6 independent multitask optimization runs and single-fidelity optimization runs. As seen in Fig. 2, the multitask algorithm exhibits an average ~10× speedup over the single-fidelity approach.

Overall, the simulation with the highest score corresponds to a case with $I_h = 4.70$ kA, $I_t = 4.67$ kA, $L_b = 6.57$ μm and $\Delta z_{l,h} = 52.8$ μm. This results in a total charge of 142 pC, an energy of 2.6 GeV (MED), and a narrow energy spread of 0.11 % (MAD). A view of the final longitudinal phase space of the beam, along with a snapshot of the plasma wakefields calculated by FBPIC and Wake-T, is shown in Fig. 3. As expected, differences in the plasma wake can be seen due to the different approximations made by each code.

CONCLUSION

This proof-of-principle study demonstrates that the proposed multitask approach effectively combines the output of simulation codes of different fidelity to speed up the optimization of an LPA stage by a factor of ~10. This automated process, which is able to extract useful information from simulation evaluations with reduced-model codes, enables cost-effective optimization of laser-plasma accelerators in large parameter spaces while retaining a high-fidelity result.

ACKNOWLEDGMENTS

This material is based upon work supported by the U.S. Department of Energy, Office of Science, under contract numbers DE-AC02-06CH11357 and DE-AC02-05CH11231 and by the Exascale Computing Project (17-SC-20-SC). This

ISBN: 978-3-95450-227-1 ISSN: 2673-5490 doi:10.18429/JACoW-IPAC2022-WEPOST030

research was supported in part through the Maxwell computational resources operated at DESY, Hamburg, Germany.

REFERENCES

[1] F. Albert *et al.*, "2020 roadmap on plasma accelerators," *New Journal of Physics*, vol. 23, no. 3, p. 031 101, 2021.

[2] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient global optimization of expensive black-box functions," *J Glob Optim*, vol. 13, no. 4, pp. 455–492, 1998.

[3] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning.* MIT Press, 2006, pp. I–XVIII, 1–248.

[4] S. Jalas *et al.*, "Bayesian optimization of a laser-plasma accelerator," *Phys. Rev. Lett.*, vol. 126, p. 104 801, 10 2021.

[5] R. Shalloo *et al.*, "Automation and control of laser wakefield accelerators using Bayesian optimization," *Nat. Comm.*, vol. 11, p. 1, 2020.

[6] A. Lifschitz, X. Davoine, E. Lefebvre, J. Faure, C. Rechatin, and V. Malka, "Particle-in-cell modelling of laser–plasma interaction using fourier decomposition," *J. Comput. Phys.*, vol. 228, no. 5, pp. 1803–1814, 2009.

[7] P. Mora and T. M. Antonsen Jr., "Kinetic modeling of intense, short laser pulses propagating in tenuous plasmas," *Phys. Plasmas*, vol. 4, no. 1, pp. 217–229, 1997.

[8] P. Sprangle, E. Esarey, and A. Ting, "Nonlinear interaction of intense laser pulses in plasmas," *Phys. Rev. A*, vol. 41, pp. 4463–4469, 8 1990, doi:10.1103/PhysRevA.41.4463

[9] K. Swersky, J. Snoek, and R. P. Adams, "Multi-task bayesian optimization," in *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[10] B. Letham and E. Bakshy, "Bayesian optimization for policy search via online-offline experimentation," *J. Mach. Learn. Res.*, vol. 20, no. 145, pp. 1–30, 2019.

[11] F. Irshad, S. Karsch, and A. Döpp, *Expected hypervolume improvement for simultaneous multi-objective and multi-fidelity optimization*, 2021.

[12] R. Lehe, M. Kirchen, I. A. Andriyash, B. B. Godfrey, and J.-L. Vay, "A spectral, quasi-cylindrical and dispersion-free particle-in-cell algorithm," *Comput. Phys. Commun.*, vol. 203, pp. 66–82, 2016.

[13] A. Ferran Pousa, R. Assmann, and A. Martinez de la Ossa, "Wake-T: A fast particle tracking code for plasma-based accelerators," *J. Phys. Conf. Ser.*, vol. 1350, no. 1, p. 012 056, 2019.

[14] P. Baxevanis and G. Stupakov, "Novel fast simulation technique for axisymmetric plasma wakefield acceleration configurations in the blowout regime," *Phys. Rev. Accel. Beams*, vol. 21, p. 071 301, 7 2018.

[15] C. Benedetti, C. B. Schroeder, C. G. R. Geddes, E. Esarey, and W. P. Leemans, "An accurate and efficient laser-envelope solver for the modeling of laser-plasma accelerators," *Plasma Phys. Control. Fusion*, vol. 60, no. 1, p. 014 002, 2017.

[16] S. van der Meer, "Improving the power efficiency of the plasma wakefield accelerator," CERN, Tech. Rep. CERN-PS-85-65-AA. CLIC-Note-3, 1985.

[17] M. Tzoufras *et al.*, "Beam loading in the nonlinear regime of plasma-based acceleration," *Phys. Rev. Lett.*, vol. 101, p. 145 002, 14 2008.

[18] K. V. Lotov, "Efficient operating mode of the plasma wakefield accelerator," *Phys. Plasmas*, vol. 12, no. 5, p. 053 105, 2005.

[19] M. Kirchen *et al.*, "Optimal beam loading in a laser-plasma accelerator," *Phys. Rev. Lett.*, vol. 126, p. 174 801, 17 2021.

[20] J.-L. Vay, "Noninvariance of space- and time-scale ranges under a Lorentz transformation and the implications for the study of relativistic interactions," *Phys. Rev. Lett.*, vol. 98, p. 130 405, 13 2007.

[21] J.-L. Vay *et al.*, "Modeling of 10 GeV-1 TeV laser-plasma accelerators using Lorentz boosted simulations," *Phys. Plasmas*, vol. 18, no. 12, p. 123 103, 2011.

[22] E. Bakshy *et al.*, "AE: A domain-agnostic platform for adaptive experimentation," in *32nd conference on Neural Information Processing Systems*, 2018.

[23] S. Hudson, J. Larson, J.-L. Navarro, and S. M. Wild, "libEnsemble: A library to coordinate the concurrent evaluation of dynamic ensembles of calculations," *IEEE Trans. Parallel Distrib. Syst.*, vol. 33, no. 4, pp. 977–988, 2022, doi:10.1109/TPDS.2021.3082815

[24] A. B. Owen, "Scrambling Sobol' and Niederreiter–Xing points," *J. Complex.*, vol. 14, no. 4, pp. 466–489, 1998.