Computer Physics Communications 265 (2021) 108026

Contents lists available at ScienceDirect

Computer Physics Communications

www.elsevier.com/locate/cpc

# A parametrized Kalman filter for fast track fitting at LHCb[☆]

P. Billoir[a], M. De Cian[b,*], P.A. Günther[c], S. Stemmle[c]

[a] LPNHE, Sorbonne Université, Paris Diderot Sorbonne Paris Cité, CNRS/IN2P3, Paris, France
[b] Institute of Physics, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[c] Physikalisches Institut, Ruprecht-Karls-Universität Heidelberg, Heidelberg, Germany

ARTICLE INFO

*Article history:*
Received 5 February 2021
Accepted 4 May 2021
Available online 7 May 2021

*Keywords:*
Kalman filter
Track reconstruction
LHCb experiment

ABSTRACT

We present an alternative implementation of the Kalman filter employed for track fitting within the LHCb experiment. It uses simple parametrizations for the extrapolation of particle trajectories in the field of the LHCb dipole magnet and for the effects of multiple scattering in the detector material. A speedup of more than a factor of four is achieved while maintaining the quality of the estimated track quantities. This Kalman filter implementation could be used in the purely software-based trigger of the LHCb upgrade.

## 1. Introduction

The LHCb experiment is a dedicated heavy flavour physics experiment at the LHC focusing on the study of hadrons containing $b$ and $c$ quarks [1]. Due to the high luminosity at the LHC and the high proton-proton interaction cross section, a sophisticated trigger system is needed to reduce the rate of collisions saved for offline analysis. During Runs 1 and 2 of the LHC, this trigger system consisted of a hardware stage, reducing the rate from 40 MHz to 1 MHz, followed by a two-stage software trigger. In the latter, the full tracking system was read out and a partial (first stage) and full (second stage) event reconstruction were performed [2]. Both software stages included a fit of selected track candidates using a Kalman filter to extract their parameters and to reject fake tracks. In addition, the software trigger allowed an online calibration and alignment of the detector [3].

During Run 3 of the LHC, LHCb will be provided with a factor five higher luminosity compared to Run 2. In this scope, most of the subdetectors are currently being replaced or upgraded [4–7] and a new trigger strategy has been developed [8]. The hardware trigger will be removed and a two-stage, fully software-based trigger will process the full 30 MHz[1] of bunch-crossing rate. In the first stage, tracks with a high transverse momentum ($p_{\mathrm{T}}$) and primary vertices will be reconstructed. These objects are used to select events with displaced topologies typical for $b$-hadron and $c$-hadron decays, and to select high-$p_{\mathrm{T}}$ objects from decays of heavy vector bosons. In the second stage, a full event reconstruction will be performed, without any requirement on the $p_{\mathrm{T}}$ and including particle identification. A large number of exclusive and several universal event selections based on the decay topology will be applied.

In LHCb, track reconstruction is split into a *pattern recognition* and a *Kalman filtering* [9,10] stage. During pattern recognition, sets in each subdetector are constructed from signals that potentially result from the passage of a single charged particle. Simple parametrizations are used throughout this procedure as it is only concerned with finding the right sets of signals and not to provide the best estimate of the track parameters. During the filtering stage, an estimate for the track parameters is calculated, and fake tracks are rejected. Given that the output of the filtering stage is used for physics selections the best possible precision needs to be achieved, hence an (extended) Kalman filter is used for track fitting. Ideally, Kalman filtering of the track candidates is already performed during the first trigger stage. However, the Kalman filter which was used during Run 1 and 2 in LHCb, in the following called *default Kalman*, is significantly too slow. It relies on lookup tables for the magnetic field and the material distribution of the detector [11], so-called *maps*. In addition it uses Runge-Kutta methods to solve the differential equations necessary to propagate the particle through the regions with an inhomogeneous magnetic field. Accessing the values in the lookup table and solving the differential equations are time consuming and prohibit the usage of the current Kalman filter in the first stage of the upgraded trigger system. This conclusion is independent of the choice of computing architecture (CPU or GPU) which is used for the first trigger stage.

[☆] The review of this paper was arranged by Prof. Z. Was.

[*] Corresponding author.
*E-mail address:* michel.de.cian@cern.ch (M. De Cian).

[1] The nominal bunch-crossing frequency of the LHC is 40 MHz, however empty and non-colliding bunches reduce this to a collision frequency of 30 MHz at LHCb.

https://doi.org/10.1016/j.cpc.2021.108026
0010-4655/

In this paper, a fully parametrized version of the Kalman filter in LHCb, called *parametrized Kalman*, is presented. It obtains precise values of track parameters and track quality variables, while relying on neither computationally costly extrapolation methods nor material or magnetic field maps.

## 2. Detector and simulation

The LHCb detector [1] is a single-arm forward spectrometer covering the pseudorapidity range $2 < \eta < 5$. Its Run 3 configuration includes a high-precision tracking system consisting of a silicon-pixel vertex detector surrounding the *pp* interaction region [5] (VELO), a large-area silicon-strip detector (Upstream Tracker (UT)) [7] located upstream of a dipole magnet with a bending power of about 4 Tm [12], and three stations of scintillating-fibre detectors (SciFi) [7] placed downstream of the magnet. Different types of charged hadrons are distinguished using information from two ring-imaging Cherenkov detectors [13,6]. Photons, electrons and hadrons are identified by a calorimeter system consisting of an electromagnetic and a hadronic calorimeter [14,6]. Muons are identified by a system composed of alternating layers of iron and multiwire proportional chambers [15,6].

Given the lack of collision data at this point for Run 3, simulation is required to model the effects of the detector response, the detector acceptance and the imposed selection requirements. In the simulation, *pp* collisions are generated using Pythia [16] with a specific LHCb configuration [17]. Decays of unstable particles are described by EvtGen [18], in which final-state radiation is generated using Photos [19]. The interaction of the generated particles with the detector, and its response, are implemented using the Geant4 toolkit [20] as described in Ref. [21].

## 3. Principles

In the following, the Kalman filter formalism and its application in the LHCb track reconstruction is outlined. During Kalman filtering, the information from measurements at detector planes is successively combined to obtain optimal estimates of the track parameters. The track is represented as a set of states at fixed $z$-positions,[2] which are typically detector layers. Each of these states is given by $\boldsymbol{x} = (x, y, t_x, t_y, \frac{q}{p})$ and the corresponding covariance matrix $\boldsymbol{P}$, where $t_x$ and $t_y$ are the slopes with respect to the $z$ axis, $q$ the charge of the particle in units of the electron charge and $p$ its absolute momentum.

The Kalman filter procedure needs an estimate of a state as a starting point. Filtering is then a repeated application of two steps. Firstly, the current state is extrapolated to the next detector layer, and secondly, the extrapolated state is updated using the measurement in this layer. If the track has no associated measurement in this layer, the update step is omitted. These steps can be formalized as follows: given the state $(\boldsymbol{x}_{k-1|k-1}, \boldsymbol{P}_{k-1|k-1})$ at position $z_{k-1}$, the extrapolated state $(\boldsymbol{x}_{k|k-1}, \boldsymbol{P}_{k|k-1})$ at position $z_k$ is given by

$$\boldsymbol{x}_{k|k-1} = \boldsymbol{f}_k(\boldsymbol{x}_{k-1|k-1}), \tag{1}$$

$$\boldsymbol{P}_{k|k-1} = \boldsymbol{F}_k \boldsymbol{P}_{k-1|k-1} \boldsymbol{F}_k^T + \boldsymbol{Q}_k, \tag{2}$$

where the extrapolation function $\boldsymbol{f}_k(\boldsymbol{x})$ is given by five individual mappings $\boldsymbol{f}_k = (f_k^x, f_k^y, f_k^{t_x}, f_k^{t_y}, f_k^{\frac{q}{p}})$. This leads to the transport matrix $\boldsymbol{F}_k$ as

$$F_k^{ij} = \frac{\partial f_k^i}{\partial x_j}. \tag{3}$$

The noise matrix $\boldsymbol{Q}_k$ accounts for uncertainties of the extrapolation, *e.g.* due to scattering at the material of the detector layers or the material in between.

The extrapolated state is then combined with the measurement $\boldsymbol{m}_k$ in the respective detector layer to obtain the new state estimate at the position $\boldsymbol{z}_k$, $\boldsymbol{x}_{k|k}$ and $\boldsymbol{P}_{k|k}$, using the following steps:

$$\boldsymbol{r}_k = \boldsymbol{m}_k - \boldsymbol{H}_k \boldsymbol{x}_{k|k-1}, \tag{4}$$

$$\boldsymbol{S}_k = \boldsymbol{H}_k \boldsymbol{P}_{k|k-1} \boldsymbol{H}_k^T + \boldsymbol{R}_k, \tag{5}$$

$$\boldsymbol{K}_k = \boldsymbol{P}_{k|k-1} \boldsymbol{H}_k^T \boldsymbol{S}_k^{-1}, \tag{6}$$

$$\boldsymbol{x}_{k|k} = \boldsymbol{x}_{k|k-1} + \boldsymbol{K}_k \boldsymbol{r}_k, \tag{7}$$

$$\boldsymbol{P}_{k|k} = (\boldsymbol{1} - \boldsymbol{K}_k \boldsymbol{H}_k) \boldsymbol{P}_{k|k-1}. \tag{8}$$

Here $\boldsymbol{H}_k$ projects the estimated state vector to the measurement space in order to allow a calculation of the residual $\boldsymbol{r}_k$. The covariance matrix of this residual is given by $\boldsymbol{S}_k$ and is combined with the covariance matrix of the state to obtain the Kalman gain $\boldsymbol{K}_k$. The latter defines then how the estimated state is modified by the residual. The variance of the residual is given by $\boldsymbol{R}_k$.

Starting at the most upstream measurement, the measurements are successively added and the track parameters updated until the last detector layer is reached. The same procedure is repeated starting at the most downstream measurement and successively including more upstream measurements. This yields two sets of states at every measurement position, which can be combined to obtain the respective optimal state.

The quality of a track can be estimated by its $\chi^2_{\text{track}}$ value. The value at each measurement is given by:

$$\chi^2_k = \chi^2_{k-1} + \boldsymbol{r}_k^T \boldsymbol{P}_{k|k}^{-1} \boldsymbol{r}_k, \tag{9}$$

and $\chi^2_{\text{track}}$ is then simply $\chi^2_k$ after all measurements have been added using the combined, optimal states.

The optimal state estimates and the measurement information can also be used to remove measurements that show a large separation from the fitted trajectory by having a large contribution to the $\chi^2_{\text{track}}$ value. They are therefore likely to be wrongly associated to the respective track, and are so-called *outliers*. Once an outlier is removed, all Kalman filter steps are performed again. This procedure can be repeated until the maximum allowed number of outliers are removed, or no more outliers are present.

The above formalism is also the basis of the Kalman filter that is currently used for track fitting in the LHCb experiment. The extrapolation functions $\boldsymbol{f}_k$ are based on maps of the magnetic field along the trajectory and numerical models for the extrapolations. Their complexities range up to a fifth-order Runge-Kutta method. The noise matrices $\boldsymbol{Q}_{\boldsymbol{k}}$ are obtained by a dedicated model for the multiple scattering and a map of the material traversed by the particle.

In the parametrized Kalman filter presented in this paper, these two costly steps are replaced by simple parametrizations. The extrapolation functions $\boldsymbol{f}_k$ are given by analytic expressions that allow a fast evaluation and calculation of the derivatives in Equation (3). The noise matrices $\boldsymbol{Q}_k$ depend on the momentum of the particle and are parametrized by a few parameters per extrapolation step.

An important difference with respect to the default Kalman filter is the treatment of energy loss due to the interaction with the detector material. While the multiple scattering is taken directly into account, the energy loss is not part of the extrapolation functions $\boldsymbol{f}_k$, *i.e.* $f_k^{\frac{q}{p}}$ is the unity transformation. This shortcoming

[2] The detector coordinate system is chosen such that the $z$-axis is parallel to the beam line and charged particles are deflected in the direction of the $x$-axis.

is compensated by choosing the momentum of the state vectors to represent the momentum at the moment of production of the particle. Thereby, the extrapolation functions also take this initial momentum as input and thus indirectly take into account all energy loss that happened on average up to the respective detector layer. The only caveat being that $\frac{q}{p}$ after the filtering is only the best representation of the true value at the production point of the particle.

## 4. Parametrizations

Depending on the strength of the magnetic field and the typical distance between detector layers, different empirical analytical functions for the extrapolation are used.

Inside the VELO, where the magnetic field is very weak, these functions and the noise matrix are given by:

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{pmatrix} f^{x}(\boldsymbol{x}) \\ f^{y}(\boldsymbol{x}) \\ f^{t_x}(\boldsymbol{x}) \\ f^{t_y}(\boldsymbol{x}) \\ f^{\frac{q}{p}}(\boldsymbol{x}) \end{pmatrix} = \begin{pmatrix} x + 0.5[t_x + f^{t_x}(\boldsymbol{x})]\Delta z \\ y + t_y \Delta z \\ t_x + p_0^{\mathrm{V}} \frac{q}{p}(z_0 + p_1^{\mathrm{V}})\Delta z \\ t_y \\ \frac{q}{p} \end{pmatrix} \tag{10}$$

and

$$\boldsymbol{Q} = \begin{pmatrix} \left(\tilde{p}_1^{\mathrm{V}}\Delta z\right)^2 Q^{t_x t_x} & 0 & \tilde{p}_2^{\mathrm{V}}\sqrt{Q^{xx}Q^{t_x t_x}} & 0 & 0 \\ 0 & \left(\tilde{p}_1^{\mathrm{V}}\Delta z\right)^2 Q^{t_y t_y} & 0 & \tilde{p}_3^{\mathrm{V}}\sqrt{Q^{yy}Q^{t_y t_y}} & 0 \\ \tilde{p}_2^{\mathrm{V}}\sqrt{Q^{xx}Q^{t_x t_x}} & 0 & \left(\tilde{p}_0^{\mathrm{V}}\left|\frac{q}{p}\right|\right)^2 & 0 & 0 \\ 0 & \tilde{p}_3^{\mathrm{V}}\sqrt{Q^{yy}Q^{t_y t_y}} & 0 & \left(\tilde{p}_0^{\mathrm{V}}\left|\frac{q}{p}\right|\right)^2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{11}$$

where $\Delta z$ is the extrapolation distance along the $z$-direction and $z_0$ the initial or final $z$ coordinate for a downstream or upstream extrapolation, respectively. The parameters $p_0^{\mathrm{V}}$, $p_1^{\mathrm{V}}$ and $\tilde{p}_0^{\mathrm{V}}$ to $\tilde{p}_3^{\mathrm{V}}$ are the same for all upstream and downstream extrapolations inside the VELO. They are determined using simulated $B_s^0 \to \phi\phi$ decays within the LHCb software framework, where $\phi \to K^+K^-$. This simulated sample allows to create a dataset $D$, containing pairs of states representing two consecutive measurements of one track inside the VELO. In addition to the true state parameters obtained from the simulation, also an extrapolation of each state to the $z$ position of the respective other state is included in the dataset. Such extrapolation is based on the default extrapolation algorithm in LHCb [11]. This dataset allows tuning the parameters employing a minimization of the following likelihood-inspired function:

$$\prod_D \left[\mathcal{G}\left(f^s(\boldsymbol{x_1}) - \boldsymbol{x_2^s}, \sqrt{Q^{ss}}\right) + c\right]. \tag{12}$$

Here, $\mathcal{G}(x, \sigma_x)$ is a normalized Gaussian distribution centered around 0 with width $\sigma_x$. The two states of each dataset entry are represented by $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and the variable $s$ is one of the state variables, $s \in \{x, t_x, y, t_y\}$. The positive empirical constant $c$ is chosen to be small with respect to the amplitude of the Gaussian function and softens the impact of outliers.

In a first step, the extrapolation functions $f^x$ to $f^{t_x}$ are tuned individually, taking into account that $f^x$ depends on the previously determined parameters for $f^{t_x}$. These tuning minimizations employ the state vector $\boldsymbol{x}_2$ that is obtained by the extrapolation of the state vector $\boldsymbol{x}_1$. This choice improves the precision of the parametrized extrapolation, by removing the effect of multiple scattering that would be present if instead the true state was chosen for $\boldsymbol{x}_2$.

In a second step, the parameters of the extrapolation functions are fixed, and a minimization of the following function is performed:

$$\prod_D \Big[\mathcal{G}_2\Big(f^d(\boldsymbol{x_1}) - \boldsymbol{x_2^d}, f^{t_d}(\boldsymbol{x_1}) - \boldsymbol{x_2^{t_d}}, \sqrt{Q^{dd}}, \sqrt{Q^{t_d t_d}}, Q^{d t_d}/\sqrt{Q^{dd}Q^{t_d t_d}}\Big) + c\Big]. \tag{13}$$

Here, $\mathcal{G}_2(x, y, \sigma_y, \sigma_y, \rho)$ is a normalized two-dimensional Gaussian distribution centered around 0 with widths $\sigma_x$ and $\sigma_y$ and a correlation factor $\rho$. The variable $d$ is either $x$ or $y$. In this minimization, the true state vector $\boldsymbol{x}_2$ is used in order to get the correct estimate of the parameters for the respective elements of the noise matrix $\boldsymbol{Q}$.

Inside the UT and the SciFi detector stations, the magnetic field is significantly stronger than inside the VELO and higher order terms are needed for the extrapolation functions:

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{pmatrix} x + \left[p_3^{\mathrm{T}} t_x + (1 - p_3^{\mathrm{T}}) f^{t_x}(\boldsymbol{x})\right]\Delta z \\ y + \left[p_5^{\mathrm{T}} t_y + (1 - p_5^{\mathrm{T}}) f^{t_y}(\boldsymbol{x})\right]\Delta z \\ t_x + \left[p_0^{\mathrm{T}}\frac{q}{p} + p_1^{\mathrm{T}}(\frac{q}{p})^3 + p_2^{\mathrm{T}} y^2 \frac{q}{p}\right]\Delta z \\ t_y + p_4^{\mathrm{T}}\frac{q}{p} t_x \frac{y}{|y|} \\ \frac{q}{p} \end{pmatrix}. \tag{14}$$

The noise matrix is given in full analogy to Equation (11) with the parameters $\tilde{p}_0^{\mathrm{T}}$ to $\tilde{p}_3^{\mathrm{T}}$, where T either stands for the UT or the SciFi detector. These parameters and the parameters $p_0^{\mathrm{T}}$ to $p_4^{\mathrm{T}}$ are individually determined on simulation for every step from one detector layer to the next and for the upstream and downstream extrapolation separately. The same strategy as for the tuning of the parameters related to the extrapolation inside the VELO is followed.

For the long extrapolations between the different tracking subdetectors, more sophisticated parametrizations are necessary. In the case of the step between the VELO and the UT, where the magnetic field is still weak, the extrapolation is based on two equations. The first describes the change in momentum along the $x$-direction of the particle:

$$\Delta p_x = p\left(\frac{t_{x,\mathrm{UT}}}{\sqrt{1 + t_{x,\mathrm{UT}}^2 + t_{y,\mathrm{UT}}^2}} - \frac{t_{x,\mathrm{V}}}{\sqrt{1 + t_{x,\mathrm{V}}^2 + t_{y,\mathrm{V}}^2}}\right) = q\int (\mathrm{d}\boldsymbol{l} \times \boldsymbol{B})_x, \tag{15}$$

where $t_{x/y,\mathrm{UT}}$ and $t_{x/y,\mathrm{V}}$ are the state variables at the first UT detector layer and the last measurement inside the VELO, respectively. The right hand side of the equation consists of an integral of the magnetic field along the trajectory of the particle. Note that the integral expression is simply a parameter which was fitted for on the dataset. The second ingredient for the extrapolation is to model the effect of the magnetic field as a single kink of the trajectory at a certain $z$-position $z_{\mathrm{mag}}$ between the VELO and the UT:

$$x_{\mathrm{UT}} = x_{\mathrm{V}} + (z_{\mathrm{mag}} - z_{\mathrm{V}})t_{x,\mathrm{V}} + (z_{\mathrm{UT}} - z_{\mathrm{mag}})t_{x,\mathrm{UT}}, \tag{16}$$

where $z_{\mathrm{V}}$ and $z_{\mathrm{UT}}$ are the positions of the states inside the VELO and the UT, respectively.

Equation (15) can be solved for $t_{x,\mathrm{UT}}$ and Equation (16) is then employed to get an expression for $x_{\mathrm{UT}}$. The unknowns in these expressions are parametrized as a function of the state variables inside the VELO:

$$t_{y,\mathrm{UT}} = t_{y,\mathrm{V}} + p_0^{\mathrm{S}} \frac{q}{p} t_{x,\mathrm{V}} \frac{y_{\mathrm{V}}}{|y_{\mathrm{V}}|} \tag{17}$$

$$\int (\mathrm{d}\boldsymbol{l} \times \boldsymbol{B})_x = p_1^{\mathrm{S}} + p_2^{\mathrm{S}} z_{\mathrm{V}} + p_3^{\mathrm{S}} t_{y,\mathrm{V}}^2 \tag{18}$$

$$z_{\mathrm{mag}} = p_4^{\mathrm{S}} + p_5^{\mathrm{S}} z_{\mathrm{V}} + p_6^{\mathrm{S}} z_{\mathrm{V}}^2 + p_7^{\mathrm{S}} t_{y,\mathrm{V}}^2. \tag{19}$$

In addition, the $y$-position of the extrapolated state is given by:

$$y_{\mathrm{UT}} = y_{\mathrm{V}} + \left[ p_8^{\mathrm{S}} t_{y,\mathrm{V}} + (1 - p_8^{\mathrm{S}}) t_{y,\mathrm{UT}} \right] \Delta z, \tag{20}$$

where $\Delta z$ is defined as the difference between $z_{\mathrm{UT}}$ and $z_{\mathrm{V}}$. The noise matrix is defined in analogy to Equation (11) with the parameters $\tilde{p}_0^{\mathrm{S}}$ to $\tilde{p}_3^{\mathrm{S}}$. These parameters and the parameters $p_0^{\mathrm{S}}$ to $p_8^{\mathrm{S}}$ are individually determined for the upstream and downstream extrapolation. The same strategy as for the tuning of the parameters related to the extrapolation inside the VELO is followed.

The extrapolation from the UT to the SciFi detector is more delicate because it is done over a distance of more than 5 meters through a strong magnetic field. Moreover, this field is far from uniform - in particular, it varies rapidly in the upper and lower regions, close to the magnet yoke. To ensure a good quality of the global track fit, the error on the extrapolation should be well below the other sources of error, mainly multiple scattering. The chosen solution is an expansion of the magnetic deviation in powers of $q/p$. The parametrization aims at giving good precision for charged particles used in physics analyses, that is for trajectories which roughly come from the origin.

To do so, the *ideal* direction $(t_x^0, t_y^0)$ as the one of a particle of charge $q$, momentum $p$, starting from the origin and hitting the UT detector layer in a given point $(x, y)$ is defined. As a good approximation, we can take $t_x^0 = x/z + \mathcal{B}q/p$, $t_y^0 = y/z$, where $\mathcal{B}$ is proportional to the integrated field between the origin and the UT. The deviations from the ideal direction, $\delta t_x = t_x - t_x^0$, $\delta t_y = t_y - t_y^0$, are small, so only a first order expansion in $\delta t_x, \delta t_y$ is considered. Corrections of higher order would be negligible compared to multiple scattering errors.

Finally, a polynomial expansion in $q/p$ for the ideal direction is built, and a correction in $\delta t_x, \delta t_y$ with coefficients which are themselves polynomials of $q/p$ is added:

$$f^x(\boldsymbol{x}) = x + t_x \Delta z + \sum_{k=1}^{K_1} A_k^x(x, y) \left(\frac{q}{p}\right)^k + \sum_{k=1}^{K_2} \left(B_k^x(x, y)\, \delta t_x + C_k^x(x, y)\, \delta t_y\right) \left(\frac{q}{p}\right)^k, \tag{21}$$

where the first two terms are the straight line extrapolation, and the next ones the curvature correction. Similar expressions are used for the other state parameters $f^y(\boldsymbol{x})$, $f^{t_x}(\boldsymbol{x})$, $f^{t_y}(\boldsymbol{x})$. The degrees of expansion $K_1$ and $K_2$ are tuned for each parameter to obtain the required precision. In practice $K_1 = 9$, $K_2 = 7$ for $f^x$ and $f^{t_x}$ and $K_1 = 7$, $K_2 = 5$ for $f^y$ and $f^{t_y}$ are used.

The dependence on $x, y$ of the coefficients $A_k^u$, $B_k^u$, $C_k^u$, with $u = x, y, t_x, t_y$, is described through a tabulation on a grid of 50 $\times$ 50 points regularly spaced on the rectangle defined by $|x/z| \leq 0.25$, $|y/z| \leq 0.25$, by steps $\Delta X$, $\Delta Y$. In order to avoid a systematic convexity bias of a bilinear interpolation, the values at $x, y$ are computed by a quadratic interpolation between the tabulated values at the six closest points on the grid: if $(X, Y)$ is the closest one, these values are: $F_{00} = (X, Y)$, $F_{+0} = F(X + \Delta X, Y)$, $F_{-0} = F(X - \Delta X, Y)$, $F_{0+} = F(X, Y + \Delta Y)$, $F_{0-} = F(X, Y - \Delta Y)$, and $F_{\varepsilon_x \varepsilon_y} = F(X + \varepsilon_x \Delta X, Y + \varepsilon_y \Delta Y)$, where $\varepsilon_x$ and $\varepsilon_y$ are the signs of $\xi = (x - X)/\Delta X$ and $\psi = (y - Y)/\Delta Y$, respectively. With these notations the interpolation formula for a quantity $F$ is given by:

$$F(x, y) = F_{00} + F_d\, \xi \psi + \big((F_{+0} - F_{-0})\, \xi + (F_{0+} - F_{0-})\, \psi + (F_{+0} + F_{-0} - 2F_{00})\, \xi^2 + (F_{0+} + F_{0-} - 2F_{00})\, \psi^2\big)/2 \tag{22}$$

$$\text{with} \quad F_d = \varepsilon_x \varepsilon_y (F_{00} + F_{\varepsilon_x \varepsilon_y} - F_{\varepsilon_x 0} - F_{0 \varepsilon_y}). \tag{23}$$

The tabulated values are obtained using the standard Runge-Kutta method of order 4, with 20 values of $q/p$ in the range $(-1/p_{min}, 1/p_{min})$, with $p_{min} = 3000$ MeV/$c$ and a polynomial fit in $q/p$. As a consequence, they do not give a reliable result for momenta below $p_{min}$. Another limitation is the larger errors on the edges of the acceptance, especially for $|t_y| \simeq 0.25$, where the field has strong spatial variations.

## 5. Performance

A sample of simulated proton-proton collisions that include a $B_s^0 \to \phi\phi$, $\phi \to K^+K^-$ decay is used to compare the reconstruction quality of the parametrized and the default Kalman filter. The extrapolation of the most upstream state estimate to the beam line is the same in both filters and is based on a simplified material map of the detector [11]. Therefore, not the state near the beam line, but the state at the most upstream measurement is employed for the comparison of the two Kalman filters. Although only tracks with measurements in each of the subdetectors are considered for this study, this is in principle not a requirement for operating the parameterized Kalman filter

Fig. 1 compares the resolution of the momentum, the $x$-position and the slope $t_x$ as a function of the true momentum of a particle. Since the position and slope are nearly exclusively determined by the measurements in the VELO, where only a very weak magnetic field is present, the parametrizations of the parametrized Kalman filter are sufficient to obtain results comparable to the default Kalman filter in these variables. In contrast, the momentum estimate strongly depends on the extrapolations in regions with strong magnetic field. There, especially at momenta below 10 GeV/$c$, an up to 20% worse resolution is observed for the parametrized Kalman filter.

The Kalman filter does not only provide an estimate of the state parameters, but also a corresponding covariance matrix. In Fig. 2 the pull distributions of the estimated momentum, $x$-position and slope $t_x$ for the parametrized Kalman filter are shown. In all three cases, good uncertainty estimates are visible. However, in analogy to the observations made for the resolution, the pull distribution of the momentum features slightly more pronounced tails.

Besides the estimate of the state near the beam line, which is used for the reconstruction of charged particles, an important output of the Kalman filter is the fit quality described by the $\chi^2_{\mathrm{track}}$ per degrees of freedom $N_{\mathrm{dof}}$. In Fig. 3, this quantity is shown for the parametrized Kalman filter for real tracks coming from a particle and fake tracks consisting of random combinations of clusters. In addition, the real track efficiencies and fake track rejection rates are shown for both Kalman filter versions when applying upper bounds on this quantity. The parametrized Kalman filter shows a slightly worse but overall comparable performance in separating the two track classes.

The fitted tracks are combined to reconstruct $B_s^0 \to \phi\phi$ candidates. Fig. 4 shows the invariant mass distribution of candidates based on the two Kalman filter versions. A single Gaussian distribution and a first order polynomial are employed to model the signal peak and the combinatorial background, respectively. This yields nearly identical estimated mass resolutions of 12.8 MeV/$c^2$ and 12.9 MeV/$c^2$ for the default and the parametrized Kalman filter, respectively.

In order to compare the timing performance of the parametrized Kalman filter and the default Kalman filter, throughput studies on a machine with two Intel(R) Xeon(R) Silver 4214 processors

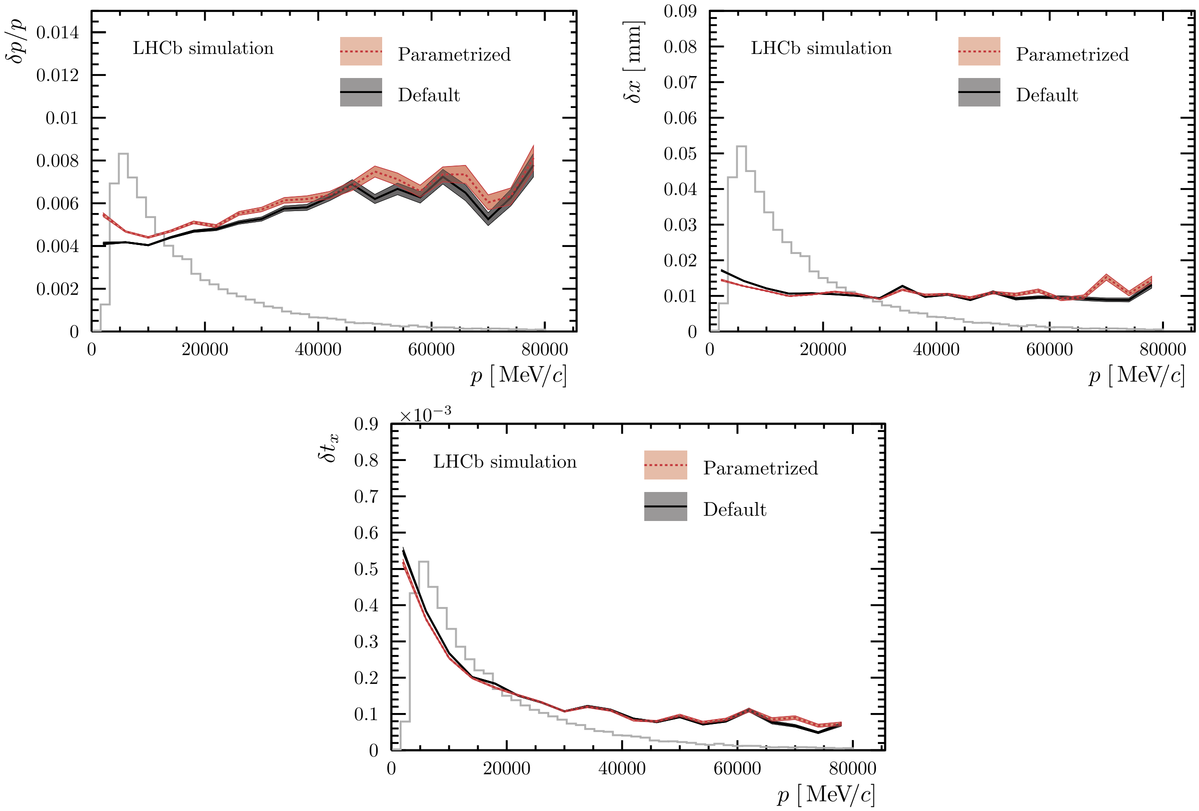

**Fig. 1.** Comparison of the resolution in simulation in (top left) momentum, (top right) $x$-position and (bottom) slope $t_x$ between the default and parametrized Kalman filter. The resolution is represented by the root mean square of the residual distribution when comparing to the true value.

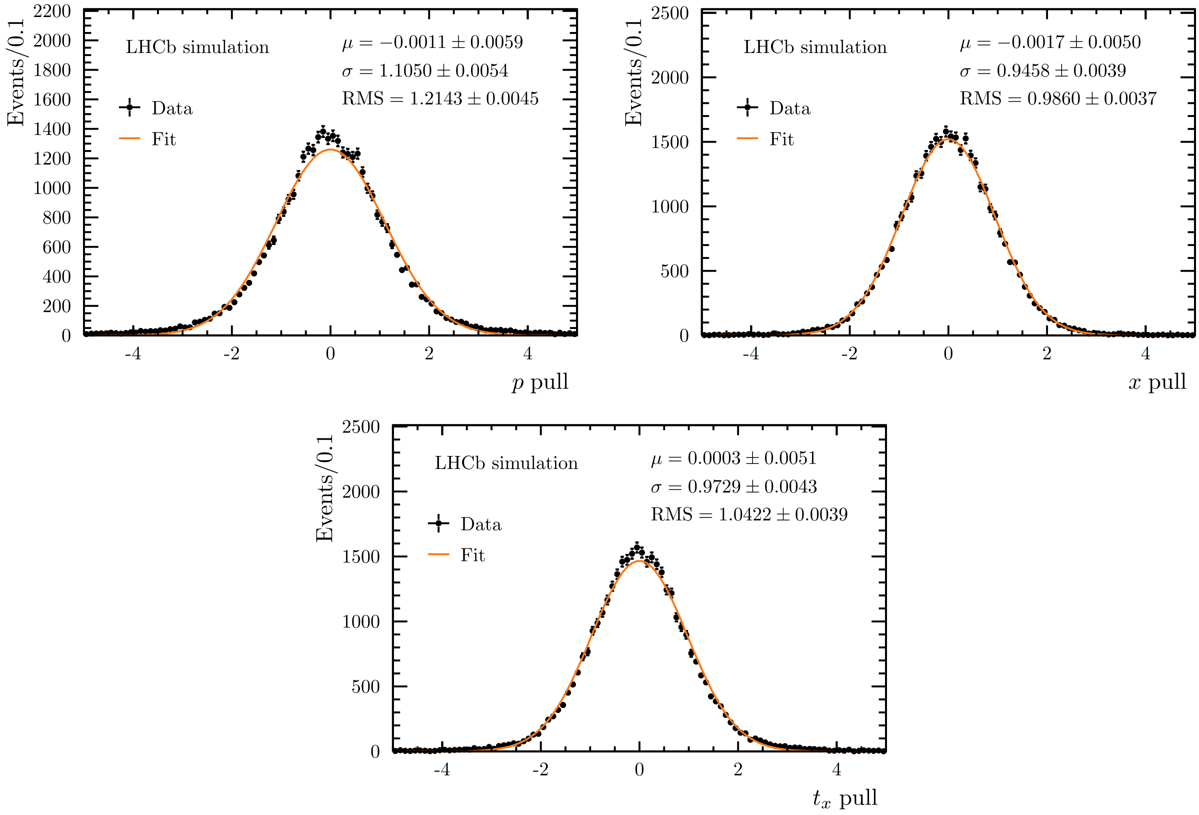

**Fig. 2.** Pull distributions of the momentum, $x$-position and slope $t_x$ estimates of the parametrized Kalman filter at the most upstream measurement. The given values correspond to the mean, width and root mean square of a Gaussian function that is fitted to the distribution.

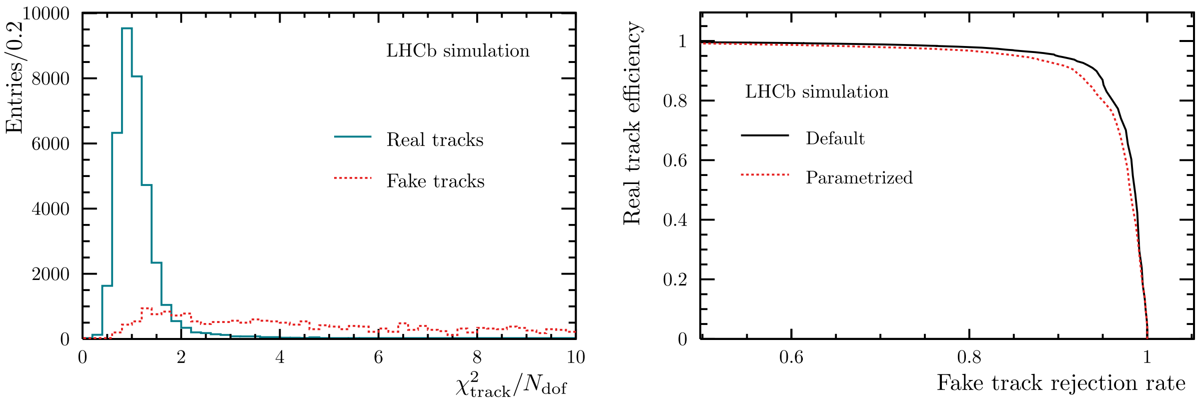

**Fig. 3.** Track quality estimate, $\chi^2_{\text{track}}/N_{\text{dof}}$, in simulation for the parametrized filter (left). Fake tracks are shown in red and real tracks in black. Real track efficiency and fake track rejection for the parametrized and default Kalman filter (right). (For interpretation of the colours in the figure(s), the reader is referred to the web version of this article.)

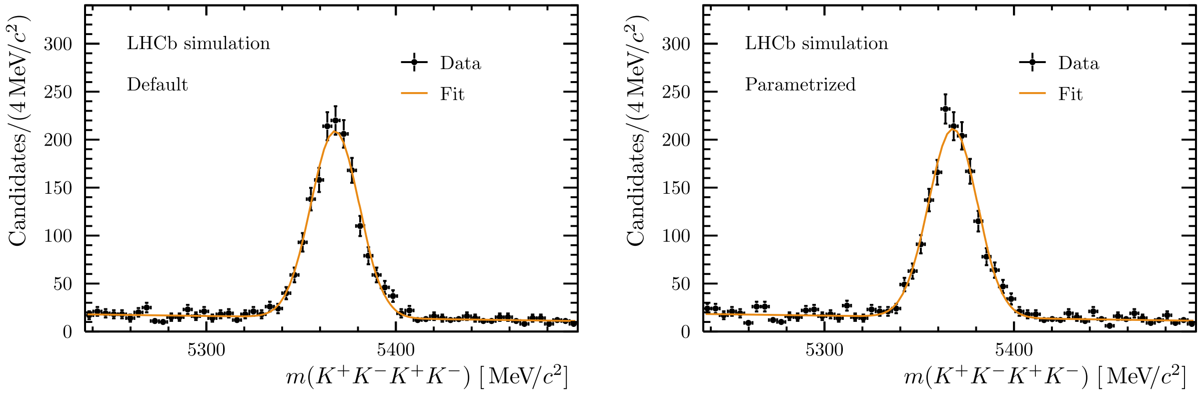

**Fig. 4.** Reconstructed $B_s^0$ mass in simulated $B_s^0 \to \phi\phi$ decays for the parametrized and the default Kalman filter. Fit projections are overlaid.

were performed. Simulated proton-proton collisions were used in order to mimic the situation of real data taking. Depending on the configuration of the outlier removal strategy, an overall speedup factor between 4 and 5.5 with respect to the default Kalman filter was achieved. The largest speedup is achieved when no iterations for the outlier removal are performed. Singling out the calculation steps of the Kalman filter, *i.e.* neglecting the part of the algorithms where the measurement information is constructed, the speedup factor is even larger and ranges from 5.7 to 10.

In the case of the parametrized Kalman filter, and singling out again the calculation step of the Kalman filter, 50% of the time is spent extrapolating the states between the detector layers. Here, the extrapolation between the UT and the SciFi constitutes the biggest component with a relative fraction of 40%. The remaining Kalman filter steps, consisting of updating the states with the cluster information and the combination of upstream and downstream filtered states, are responsible for 16% and 14% of the time spent, respectively. The extrapolation to the beam line, which is based on the default LHCb extrapolation algorithm, is responsible for the remaining 20% of the time budget.

## 6. Conclusion

We presented an alternative implementation of a Kalman filter for the LHCb experiment. Based on simple parametrizations of material effects and the extrapolation through the magnetic field of the detector, this algorithm achieves a significant speedup with respect to the current implementation, while retaining comparable quality of the track parameters. In the future, further improvements of the parametrizations might allow an even better estimate of the track parameters and a subsequent speedup. Ideas currently under discussion include for example an analytic parametrization of the $x$ and $y$ dependence of the parameters employed in the extrapolation from the UT to the SciFi detector and a better account for the limited acceptance of low momentum particles. The version presented in this document or a future implementation might therefore be well suited for the usage in the LHCb software trigger system for Run 3 of the LHC.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

The authors would like to thank the LHCb computing and simulation teams for their support and for producing the simulated LHCb samples used in the paper. We also would like to thank the LHCb RTA team for supporting this publication and reviewing the work. M. De Cian acknowledges support from the Swiss National Science Foundation grant “Probing right-handed currents in quark flavour physics”, PZ00P2_174016.

## References

[1] LHCb collaboration, A.A. Alves Jr., et al., J. Instrum. 3 (2008), S08005.
[2] R. Aaij, et al., J. Instrum. 14 (2019) P04013, arXiv:1812.10790.

[3] S. Borghi, in: Proceedings of the Vienna Conference on Instrumentation 2016, Nucl. Instrum. Methods Phys. Res., Sect. A, Accel. Spectrom. Detect. Assoc. Equip. 845 (2017) 560.
[4] LHCb collaboration, I. Bediaga, et al., Framework TDR for the LHCb Upgrade: Technical Design Report, CERN-LHCC-2012-007, 2012.
[5] LHCb collaboration, I. Bediaga, et al., LHCb VELO Upgrade Technical Design Report, CERN-LHCC-2013-021, 2013.
[6] LHCb collaboration, I. Bediaga, et al., LHCb PID Upgrade Technical Design Report, CERN-LHCC-2013-022, 2013.
[7] LHCb collaboration, A.A. Alves Jr., et al., LHCb Tracker Upgrade Technical Design Report, CERN-LHCC-2014-001, 2014.
[8] LHCb collaboration, I. Bediaga, et al., LHCb Trigger and Online Technical Design Report, CERN-LHCC-2014-016, 2014.
[9] R.E. Kalman, J. Basic Eng. 82 (1960) 35, https://asmedigitalcollection.asme.org/fluidsengineering/article-pdf/82/1/35/5518977/35_1.pdf.
[10] R. Frühwirth, Nucl. Instrum. Methods Phys. Res., Sect. A, Accel. Spectrom. Detect. Assoc. Equip. 262 (1987) 444.
[11] E. Bos, E. Rodrigues, The LHCb Track Extrapolator Tools, Tech. Rep. LHCb-2007-140. CERN-LHCb-2007-140, CERN, Geneva, 2007.
[12] LHCb collaboration, S. Amato, et al., LHCb magnet: Technical Design Report, CERN-LHCC-2000-007 , 2000.
[13] LHCb collaboration, S. Amato, et al., LHCb RICH: Technical Design Report, CERN-LHCC-2000-037, 2000.
[14] LHCb collaboration, S. Amato, et al., LHCb calorimeters: Technical Design Report, CERN-LHCC-2000-036, 2000.
[15] LHCb collaboration, P.R.B. Marinho, et al., LHCb muon system: Technical Design Report, CERN-LHCC-2001-010, 2001.
[16] T. Sjöstrand, S. Mrenna, P. Skands, Comput. Phys. Commun. 178 (2008) 852, arXiv:0710.3820;
T. Sjöstrand, S. Mrenna, P. Skands, J. High Energy Phys. 05 (2006) 026, arXiv:hep-ph/0603175.
[17] I. Belyaev, et al., J. Phys. Conf. Ser. 331 (2011) 032047.
[18] D.J. Lange, Nucl. Instrum. Methods A 462 (2001) 152.
[19] P. Golonka, Z. Was, Eur. Phys. J. C 45 (2006) 97, arXiv:hep-ph/0506026.
[20] Geant4 collaboration, J. Allison, et al., IEEE Trans. Nucl. Sci. 53 (2006) 270;
Geant4 collaboration, S. Agostinelli, et al., Nucl. Instrum. Methods A 506 (2003) 250.
[21] M. Clemencic, et al., J. Phys. Conf. Ser. 331 (2011) 032023.