Non-equilibrium spin transport in integrable spin chains: persistent currents and emergence of magnetic domains

Andrea De Luca,1 Mario Collura,1 and Jacopo De Nardis2

1The Rudolf Peierls Centre for Theoretical Physics, Oxford University, Oxford, OX1 3NP, United Kingdom
2Département de Physique, Ecole Normale Supérieure PSL Research University, CNRS, 24 rue Lhomond, 75005 Paris, France

We construct exact steady states of unitary non-equilibrium time evolution in the gapless XXZ spin-1/2 chain where integrability preserves ballistic spin transport at long-times. We characterize the quasi-local conserved quantities responsible for this feature and introduce a computationally effective way to evaluate their expectation values on generic matrix product initial states. We employ this approach to reproduce the long-time limit of local observables in all quantum quenches which explicitly break particle-hole or time-reversal symmetry. We focus on a class of initial states supporting persistent spin currents and our predictions remarkably agree with numerical simulations at long times. Furthermore, we propose a protocol for this model where interactions, even when antiferromagnetic, are responsible for the unbounded growth of a macroscopic magnetic domain.

Introduction. — One-dimensional many-body physics displays a rich landscape of exactly solvable models, which offers a perfect outpost to study the role of interaction in quantum mechanics. As a remarkable example, integrable spin chains have provided in the past decades, numerous insights on the equilibrium properties of quantum magnets, then confirmed in several experiments [1]. On the other hand, the study of non-equilibrium properties has a much shorter history, as only recently isolated quantum systems have been successfully prepared in experiments and studied under their unitary time evolution [2]. In these settings, it was possible to observe with high precision transport processes in many-body 1D systems [3] and relaxation to steady states after a non-equilibrium time evolution [4].

From the theoretical point of view, the focus has been on understanding how a unitarily evolving system can relax to a stationary state [5] and, if this is the case, how to predict the stationary values of local observables. The XXZ spin-1/2 chain defined by the Hamiltonian (1) represents a perfect setting where to understand the role of interactions and local symmetries in the non-equilibrium time evolution [7–11]. In particular, in one dimension, two main non-equilibrium protocols have been considered: a) the so-called quantum quench, when starting from a pure state $|\Psi_0\rangle$, the system is let unitarily evolve under its Hamiltonian $H = -i\partial_t |\Psi_0\rangle$ [8] and b) the open Markovian case, when the edges of the chain are kept in contact with two spin baths [13] (a similar protocol has also recently been realized in experiments [16]).

The study of the case a) has led to a neat formulation [10, 11] of the Generalized Gibbs Ensemble (GGE) [12]: the steady state is locally described by a density matrix of Gibbs-like form $\rho = e^{-\sum_j \beta_j Q_j}$, with $\{Q_j\}$ a set of local and quasi-local conserved quantities, namely extensive operators commuting with the global Hamiltonian $[H, Q_j] = 0$. The $\beta_j$’s are fixed requesting that the stationary expectation values of all the conserved charges coincide with the initial ones. This statement extends to any interacting model once an appropriate set $\{Q_j\}$ has been identified.

The protocol b) has shown how steady states can exhibit a genuine non-equilibrium behavior and persistent spin currents. At the origin of this peculiar behavior, lies the existence of a set of of previously unknown quasi-local conserved operators $\{Z_n\}$ which preserve “ballistic” spin transport [17]. This set of conserved operators is characteristic of the gapless phase and is responsible for the existence of a finite Drude weight [17–19], which has attracted large interest in the recent years [20]. However, it was unclear if such quasi-local operators $\{Z_n\}$ could be included into a GGE-like steady state resulting from the long-time dynamics of a close system. If this is the case, one can choose initial states $|\Psi_0\rangle$, which, by pure unitary time evolution, lead to a persistent current in the stationary state, namely $\lim_{t\to\infty} \langle J_i |\Psi_0(t)\rangle \neq 0$. Here current operators are defined by continuity equations, e.g. for the local magnetizations, $ds_i^z/dt = \langle [H, s_i^z] \rangle = J_{i-1} - J_i$.

In this work, we solve this open issue. We show that the unitary evolution of a generic state $|\Psi_0\rangle$ with the gapless XXZ Hamiltonian, leads to a GGE steady state which includes the set $\{Z_n\}$. Ignoring these quantities in the GGE produces significative discrepancies with the large-time limit of local observables, as was observed in a specific example [31]. Moreover, we present an efficient way to handle such charges in practice, which is applicable to any initial state expressed as a matrix product state. We obtain a class of non-equilibrium time evolutions exhibiting persistent spin transport at large times, beyond the linear response regime described by the equilibrium Drude weight (accessible isotopically instead with local quantum quenches [21]).

However, to our construction, non-trivial steady state can be engineered by joining two initial states with different spin-transport properties, then confirmed in several experiments [1]. On the other hand, the study of non-equilibrium properties has a much shorter history, as only recently isolated quantum systems have been prepared in experiments and studied under their unitary time evolution [2]. In these settings, it was possible to observe with high precision transport processes in many-body 1D systems [3] and relaxation to steady states after a non-equilibrium time evolution [4].

...
main inside the antiferromagnetic phase of the XXZ chain (1).

The model and classification of steady states. — We consider the XXZ spin-1/2 chain in the gapless phase \( |\Delta| < 1 \)

\[
H = \sum_{i=-\frac{1}{2} + 1}^{\frac{1}{2}} \left[ s_i^x s_{i+1}^x + s_i^y s_{i+1}^y + \Delta \left( s_i^z s_{i+1}^z - \frac{1}{4} \right) \right], \tag{1}
\]

(with \( \{ s_i^a \} \) indicating the spin-1/2 operators) and parametrize the anisotropy as \( \Delta = \cos \gamma \). The model is solvable by Bethe ansatz: every eigenstate \( \{ \lambda_k \} \) is labeled by a set of \( N \) complex “rapidities” \( \{ \lambda_k \} \) ful-
filling the Bethe equations [28]. In the thermodynamic limit \( L \to \infty \), rapidities are arranged in strings, composed of \( n_j \) rapidities, sharing the same real part \( \lambda_j^{(\alpha)} \) and equispaced imaginary parts, with gaps of size \( \gamma \). These solutions can be interpreted as bound states of magnons with different lengths \( \{ n_j \} \) and center of mass momenta \( \{ \lambda_j^{(\alpha)} \} \) with \( \alpha \) labeling all bound states with the same length. Whenever \( \gamma/\pi \) is a rational number, the total number of string types is finite. In the follow-
ing, we focus on the simplest case \( \gamma = \pi/\ell \), where one has exactly \( \ell \) bound-state types.

In the thermodynamic limit the string momenta \( \{ \lambda_j^{(\alpha)} \} \) become dense on the real line and for each string type \( j \), we can introduce the corresponding density distribution \( \rho_j(x) = \lambda_j^{(\alpha)} - \lambda_j^{(\alpha)} - 1 \). We define then a macroscopic eigenstate \( |\rho_j(\lambda)\rangle \) [23] as a set of particle dis-
tributions \( \{ \rho_j(\lambda) \} \), one for each string type. Then, the expectation value of any extensive conserved charge \( Q = \sum_{i} q_i \), with \( q_i \) the charge density, is expressed in the thermodynamic limit only in terms of the distribution of particles and its single-particle eigenvalues \( q_j(\lambda) \)

\[
\langle \rho | Q | \rho \rangle = L \sum_{j=1}^{\ell} \int_{-\infty}^{+\infty} d\lambda \rho_j(\lambda) q_j(\lambda). \tag{2}
\]

In particular, the total number of particles, i.e. the number of down spins in the system \( Q = L/2 - \sum_i s_i^z \), cor-
sponds to the case \( q_j(\lambda) = n_j \) and fixes the normalization of the densities \( \{ \rho_j(\lambda) \} \). Instead, the energy density \( e = \frac{1}{L} \langle \rho | H | \rho \rangle \) is obtained for \( q_j = -J\pi \sin(\gamma) a_j(\lambda) \), with \( a_j(\lambda) \) the scattering kernel given in (S28) of [28].

Analogously to representing the Gibbs trace with a single expectation value at thermal equilibrium [24], it is useful to associate to the canonical GGE density matrix \( \rho = e^{-\sum_j \beta_j Q_j} \), a corresponding microcanonical formulation, where expectation values are computed on a single macroscopic eigenstate [25]

\[
\frac{\text{Tr} \left( e^{-\sum_j \beta_j Q_j} \mathcal{O} \right)}{\text{Tr} \left( e^{-\sum_j \beta_j Q_j} \right)} = \langle \rho | \mathcal{O} | \rho \rangle \tag{3}
\]

for any local operator \( \mathcal{O} \). The densities \( \{ \rho_j \} \) associated to the macroscopic eigenstate in (3) are fixed matching the expectation values (2) of the conserved charges.

A complete set of charges \( \{ q_j \} \) is such that all the densities \( \{ \rho_j \} \) are univocally pinned down. A set symmetric under spin reversal was introduced in [26]: in this approach, charges were organized into a finite number of families \( \{ Q_n^{(s)} \} \) where \( s = 1/2, 1, \ldots, s \) (with \( s = \ell/2 \)), in one-to-one correspondence with irreducible representations of the \( SU(2) \) algebra \( q = e^{i\gamma} \) with \( s \) labeling the spin of the representation. Here, \( n \in \mathbb{N} \) indexes all charges for a given representation \( s \). For instance, the Hamiltonian (1) coincides with \( H = -J\pi \sin(\gamma) Q_1^{(1/2)} \).

In [10], it was found an economic way to directly relate the densities \( \{ \rho_j \} \) to the expectation values of the charges on any state \( |\Psi_0\rangle \): thanks to (3), this provides directly the stationary GGE whenever \( |\Psi_0\rangle \) is chosen as an initial state. This goes through the introduction of the generating functions defined as [10, 11]

\[
X_j(\lambda) = \frac{L}{\ell} \sum_{n=1}^{\ell} \lambda^{n-1} \langle \Psi_0 | Q_j^{(n/2)} | \Psi_0 \rangle. \tag{4}
\]

Despite their representation in terms of a complicated power series, these functions can be easily computed on any initial state of the form of a matrix product state [7]. Then, the correspondence between such generating functions and the distributions of roots can be deduced from (2) in (4) and reads

\[
\rho_j = -X_{j-1} + X_j^{[\ell]} - X_{j+1}, \quad j = 2, \ldots, \ell - 2,
\]

\[
\rho_1 = X_{[\ell]} - X_2, \quad \rho_{\ell-1} = \rho_{\ell-1} - \rho_1 = -X_{\ell-2} + X_{[\ell]}, \tag{5}
\]

where we used the notation \( f^{[\ell]} = f(\lambda + i\frac{\gamma}{2}) + \bar{f}(\lambda - i\frac{\gamma}{2}) \).

At this point, one immediately realizes that only the difference of the last two distribution is fixed by (5). In [11], the additional constraint was deduced mimicking states describing thermal equilibrium. However, its va-

validity is found to be restricted to spin-flip invariant initial
states $S|\Psi_0\rangle = \pm|\Psi_0\rangle$ with $S = \prod_i s_i^z$ the spin-flip operator on the whole chain (examples are the well-studied Neel and Majumdar-Ghosh states $|8, 9\rangle$). We now show how to extend this condition to general quench protocols.

**Generalized Gibbs Ensemble in the gapless regime.** — The set of conserved charges included in the GGE (5) is not complete in the gapless regime of the XXZ chain. Here, we provide the remaining set of charges that completely determines any steady state. In [19], it was shown that the key point to get odd conserved quantities in the periodic XXZ spin chain lies in the representation of $SU_q(2)$ corresponding to maximal spin $s = \hat{s}$. Indeed, in this further case, irreducible representations of dimension $\ell - 1$ can be obtained as a function of an additional continuous parameter $\alpha$ [27].

Similarly to what is done in (4), one can construct an $\alpha$-dependent generating function $X_{2\hat{s},\alpha}(\lambda)$, where $\alpha = 0$ reduces to the original $X_{2\hat{s}}(\lambda)$. The first difficulty one faces in exploiting, in practice, these operators is the illusive richness of new conserved quantities (an infinite family for each value of $\alpha$) which contrasts with the limited freedom left in (5), where essentially only one function remains to be fixed. This apparent paradox can be solved by showing that only $X'_{\alpha}(\lambda) \equiv \partial_\alpha X_{2\hat{s},\alpha}(\lambda)|_{\alpha=0}$ has to be added to completely fix all the density distributions. Up to a multiple of the identity, $X'(\lambda)$ generates expectation values for a family of conserved quantities $Z_n$, odd under spin flip ($SZ_nS = -Z_n$), i.e.

$$X'(\lambda) + \bar{\delta}X'(\lambda) = \frac{1}{\ell \sum_{n=1}^{\infty}} \lambda^{n-1} \langle \Psi_0 | Z_n | \Psi_0 \rangle,$$  

which, as shown in [28], can be efficiently computed for all initial states $|\Psi_0\rangle$ expressible as low bond-dimension matrix product states. The term $\bar{\delta}X'(\lambda) = \frac{-1}{2\pi \cos^2(\lambda)}$ is simply a multiple of the identity which is added for consistency with all the other $X_j(\lambda)$, all having vanishing expectation on the fully polarized state along the $z$-direction.

The second difficulty is to derive a relation between $X'(\lambda)$ and $\{\rho_j(\lambda)\}_{j=1}^\ell$, giving the missing piece in (5). Being quasi-local, the expectation value of $Z_n$ on a many-body state is additive on its particle content, in agreement with (2), but single-particle eigenvalues corresponding to each $Z_n$, i.e. the functions $q_j(\lambda)$ in (2), are not accessible with standard methods [10]. We follow therefore a different strategy: we consider eigenstates $H$ with a single magnon of rapidity $\lambda$. On these states, the expectation value of each $Z_n$ can be analytically computed, and gives access in the limit $L \to \infty$ to the required functions $q_j(\lambda)$ [28].

With this information, we can complete (5) with the additional relation

$$\rho_\ell = -\frac{1}{2} \frac{X'[\lambda]}{\ell - 1} - \frac{1}{2\gamma} \int_{-\gamma/2}^{\gamma/2} dz X'(z),$$  

where for a generic function $f$, we introduced the notation $f[z](\lambda) \equiv f(\lambda + iz)$. When the initial state $|\Psi_0\rangle$ is invariant under spin-flip inversion we have $\langle \Psi_0 | Z_n | \Psi_0 \rangle = 0$ and $X'(\lambda) = -\delta(\lambda)$. This can be shown to imply the restriction assumed in [11] (see [28]), which is therefore proved to be correct only for spin-flip invariant initial states. For all other cases, $X'(\lambda)$ is non-trivial and it is responsible for a steady spin current in the system. We consider an explicit example in the following section.

**Global quench from a dimer state.** — We introduce a simple initial state that breaks time-reversal symmetry

$$|\Psi_0^{(\phi)}\rangle = \bigotimes_{i \text{even}} \left( |\uparrow\downarrow\rangle_{i,i+1} + e^{i\phi} |\downarrow\uparrow\rangle_{i,i+1} \right)/\sqrt{2}.$$

This state is not an eigenstate of the XXZ Hamiltonian and is not invariant under spin-flip inversion as it supports a finite spin current on the even sites $\langle \Psi_0^{(\phi)} | J_{2i} | \Psi_0^{(\phi)} \rangle = \frac{\ell}{2} \sin \phi$. The evaluation of the generating functions $\{X_j(\lambda)\}_{j=1}^\ell$ and $X'(\lambda)$ is straightforward on this state and leads to a complete characterization of its steady state according to (5, 7). Denoting with $v_j(\lambda)$ the effective velocities of the bound states of type $j$ [14, 35], the steady state gives in particular the long-time limit of the spin current

$$\lim_{t \to \infty} \langle \Psi_0^{(\phi)}(t) | J_{\ell} | \Psi_0^{(\phi)}(t) \rangle = \sum_{j=1}^{\ell} \int_{-\infty}^{+\infty} d\lambda v_j(\lambda)n_j,$$

which proves to be non-zero and homogeneous on the whole chain due to the presence of the odd conserved charges $\{Z_n\}$ (see Fig. 1). Note how such non-equilibrium protocol probes the non-linear well beyond the regime accessible though the Drude weight at small $\phi$. The behavior of the stationary current as a function of $\phi$ is shown in Fig. 2. It is interesting to investigate the isotropic limit $\Delta \to 1^-$, where the stationarity of the
spin current breaks down: the charges \( \{ Z_n \} \) become non-local and the steady current vanishes, compatibly with the absence of a finite Drude weight in the isotropic chain \( [30] \). Our data are consistent, for any \( \phi \), with a critical form \( J(\Delta) \sim (1 - \Delta)^3 \).

Finally, we remark that the GGE state obtained from (5) without taking into account the charges \( \{ Z_n \} \), though correctly predicting expectation values of even local operators, would spoil the odd ones, leading for instance to a vanishing steady spin current, as pointed out in \([31]\)

**Joining two states with different spin currents.** —

We now consider the non-equilibrium steady state \([14, 15, 29, 34]\) generated by joining at time \( t = 0 \) two dimer states defined in (8) with two different phases. Taking \( |\Psi_0^{\phi_L} \rangle \) on the sites \( i \in [-L/2 + 1, \ldots, 0] \) and \( |\Psi_0^{\phi_R} \rangle \) on the sites \( i \in [1, \ldots, L/2] \), we consider

\[
|\Psi_0^{(\phi_L, \phi_R)} \rangle = |\Psi_0^{\phi_L} \rangle \otimes |\Psi_0^{\phi_R} \rangle ,
\]

and its time evolution. Using the recent developments on the inhomogeneous quantum quenches \([14, 29]\), combined with our complete characterization of both GGE states, we compute the long-time limit of the local magnetization \( \langle s^z_i \rangle \) and the local spin current \( \langle J_i \rangle \) at fixed rays \( i = c t \) in the range \( -u_L < z < u_R \) with \( u_L, v_R \) the maximal velocity of the excitations respectively of the left and right steady state \([14, 35]\). Even though the initial state has a uniform zero magnetization, we observe the formation of an expanding magnetic domain where the magnetization changes from 0 to a finite value which depends on the unbalance between the left and right steady-state spin currents (see Fig. 3).

**Conclusions.** — We introduced a complete generalized Gibbs ensemble for the XXZ spin chain that can be extended to any lattice integrable model. We showed how the set of quasi-local charges recently introduced for the gapped regime in \([26]\) is not sufficient to unambiguously determine the steady state after a quantum quench. The existence of additional quasi-local charges had been pointed out in the study of the equilibrium Drude weight \([17-19]\) but their application in the exact non-equilibrium time-evolution had remained up to now elusive. We considered the dimer state (8) as an explicit example of initial state which breaks time-reversal symmetry and displays a steady spin current in the limit \( t \to \infty \) which depends non-trivially on the initial phase-shift \( \phi \). An other interesting example amenable for simple treatment within our framework would be the quench protocol where a finite magnetic flux is suddenly switched on at \( t = 0^+ \), as recently studied on the XXZ spin chain \([31, 32]\) as well as in Chern insulators \([33]\). Our analytic approach allowed us to address the XXX limit \( \Delta \to 1 \) where the steady spin current vanishes and the behavior close to the isotropic point can be used to estimate the large-time decay.

We also showed how joining two chains generically leads to the spontaneous creation of expanding magnetic domains, whose edges are a direct measure of the quasiparticle velocities and which might have direct connections with the physics of domain growth. We remark that once the two states have been realized, only a local quench is necessary to create the junction \([34]\) making this protocol amenable for experimental tests.

**Acknowledgments.** — We are very grateful to Maurizio Fagotti for the collaboration in the early stage of this project. ADL would like to thank Dario Villamaina for useful discussion on transport in diffusive systems. This work was supported by by LabEX ENS-ICFP:ANR-10-LABX-0010/ANR-10-IDEX-0001-02 PSL* (J.D.N.), the EPSRC Quantum Matter in and out of Equilibrium Ref. EP/N01930X/1 (A.D.L.) and the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 701221 (M.C.).
De Luca, J. Viti et al, Phys. Rev. B 90, 161101(R) (2014); B. Doyon, M. Hoogeveen, and D. Bernard, J. Stat. Mech. (2014) P03002; V. Eisler, Z. Zimboras, New J. Phys. 16, 123020 (2014); M. Collura and D. Karevski, Phys. Rev. B 89 214308 (2014); M. Collura and G. Martelloni, J. Stat. Mech. (2014) P08006; A. De Luca, G. Martelloni, and J. Viti, Phys. Rev. A 91, 021603(R) (2015); B. Doyon, A. Lucas et al, J. Phys. A: Math. Theor. 48 095002 (2015); B. Doyon, Nucl. Phys. B 892, 190 (2015); J. Dubail, J.-M. Stephan, J. Viti and P. Calabrese, arXiv:1606.04401 (2016); D. Bernard, B. Doyon, arXiv:1612.05956 (2016).

35 L. Bonnes, F.H.L. Essler and A. M. Läuchli, Phys. Rev. Lett. 113, 187203 (2014).
36 C. Gómez, M. Ruiz-Altaba, and G. Sierra. Quantum groups in two-dimensional physics. Cambridge University Press, 2005.
Supplementary Material
Non-equilibrium spin transport in integrable spin chains: persistent currents and emergence of magnetic domains

Here we give additional details about the calculations presented in the letter.

- In Appendix A we introduce the whole landscape of local and quasi-local conserved quantities of the XXZ model, using all the finite \( SU_q(2) \) representations. We recover the integer spin representations, introduced in [26], and we introduce a family of conserved charges coming from a representation with non-integer spin, analogously to what done in [18, 19].

- In Appendix B we compute an analytic expression for the eigenvalues of these new charges on the Bethe eigenstates in the thermodynamic limit.

- In Appendix C we use the expressions of their eigenvalues to show the relation between the expectation values of these conserved quantities on the initial state and the distribution of rapidities \( \{ \mu_j \}_{j=1}^\gamma \) specifying the GGE steady state.

- In Appendix D we show how to evaluate the generating functions \( \{ \{ X_{2s} \}_{s=\frac{1}{2}} \}, X' \} \) on the initial state when this is a product spin state.

A. FAMILY OF CONSERVED CHARGES AND \( SU_q(2) \) REPRESENTATIONS

1. Commuting transfer matrices

We briefly summarize how the XXZ is constructed as an integrable model in the framework of the algebraic Bethe Ansatz. As mentioned in the text, we rewrite the parameter \( \Delta \) of the Hamiltonian as

\[
\Delta = \cos \gamma = \frac{q + q^{-1}}{2}, \quad q = e^{i\gamma} .
\]

We then introduce the \( L \)-matrix defined on the tensor product \( V_n \otimes V'_n \):

\[
L_{n,a}(\lambda) = \sinh(\lambda(K_n + K_n^{-1}))I_a + \cosh(\lambda(K_n - K_n^{-1}))\sigma_a^z + (q - q^{-1})(S_n^-\sigma_a^+ + S_n^+\sigma_a^-)
\]

where \( \lambda \) is the spectral parameter. Here, \( V_a = \mathbb{C}^2 \) is a associated to a spin 1/2 representation with Pauli spin operators \( \sigma_a^x, \sigma_a^y, \) while the space \( V_n \) is associated to a representation of the \( SU_q(2) \) algebra:

\[
K_n^a S_n^a = q^{a_1} S_n^a K_n, \quad [S_n^+, S_n^-] = \frac{K_n^2 - K_n^{-2}}{q - q^{-1}} .
\]

The fundamental representation of dimension 2 is easily obtained taking

\[
K_n = q^{a_1} , \quad S_n^\pm = \frac{1}{2}(\sigma_x \pm i\sigma_y) .
\]

In this case the two spaces \( V_n \sim V_{a_1} = \mathbb{C}^2 \) and \( V_a \sim V_{a_2} = \mathbb{C}^2 \) have the same dimension and the matrix \( L_{a_1,a_2}(\lambda) \) assumes a symmetric form. One can then introduce the \( R \)-matrix between them:

\[
R_{a_1,a_2}(\lambda) = L_{a_1,a_2}(\lambda + \gamma) = \begin{pmatrix}
2\sinh(\lambda + i\gamma) & 0 & 0 & 0 \\
0 & 2\sinh(\lambda) & 2i\sin\gamma & 0 \\
0 & 2i\sin\gamma & 2\sinh(\lambda) & 0 \\
0 & 0 & 0 & 2\sinh(\lambda + i\gamma)
\end{pmatrix}
\]

The matrix \( L \) and \( R \) satisfy the Yang-Baxter equation in the space \( V_a \otimes V_{a'} \otimes V_{a''} \)

\[
R_{a,a'}(\lambda - \mu)L_{a',a''}(\lambda)L_{a'',a}(\mu) = L_{a,a'}(\mu)L_{a'',a}(\lambda)R_{a,a'}(\lambda - \mu).
\]

In this equation, one interprets the spaces \( V_a, V_{a'} \) as auxiliary spaces and the space \( V_{a''} \) as the quantum space of one single physical spin. Then, it tells that the \( R \)-matrix can be used to exchange the \( L \)-matrices defined on the same
physical space. It is a direct consequence of the algebra in Eq. (S3): in particular generalizations exist for any pair of auxiliary spaces $V_n$ and $V_{n'}$, which allow to exchange $L_{n,n}$ and $L_{n,n'}$.

We then introduce the transfer matrix defined as the product of $L$ matrices acting on the Hilbert space of a chain of $L$ spin-$1/2$: $V_{a_1} \otimes \ldots \otimes V_{a_L}$

$$T_n(\lambda) = L_{n,a_1}(\lambda) L_{n,a_2}(\lambda) \cdots L_{n,a_L}(\lambda) = \bigotimes_{i=1}^{L} L_{n,a_i}(\lambda) . \quad (S7)$$

It is easy to verify that Eq. (S6) (and its generalizations to pairs of representations with $R_{n,n'}(\lambda)$) implies an analogous relation for the product of $L$-matrices, the so-called RTT relation

$$R_{n,n'}(\lambda - \mu) T_n(\lambda) T_{n'}(\mu) = T_{n'}(\mu) T_n(\lambda) R_{n,n'}(\lambda - \mu) . \quad (S8)$$

which, upon tracing over the auxiliary spaces, leads to

$$[T_n(\lambda), T_{n'}(\mu)] = 0 , \quad T_n(\lambda) = \text{tr}_n(T_n(\lambda)) . \quad (S9)$$

In this way we obtain several families of commuting operators in correspondence of each representation of the $SU_q(2)$-algebra in Eq. (S3).

2. Representation of $SU_q(2)$

Finite dimensional representations of the algebra in (S3) are known and we refer to [36] for a thorough discussion. For generic values of $q$, there is a one-to-one correspondence with the representations of $SU(2)$, labeled by the value of the spin $s$. We already showed in Eq. (S4) how a two-dimensional representation is triavilly obtained from Pauli matrices. In general, the representation of spin $s$ of the undeformed $SU(2)$ has the form

$$S^z_s |m\rangle = m |m\rangle , S^\pm_s |m\rangle = \sqrt{(s+1 \pm m)(s \mp m)} |m \pm 1\rangle . \quad (S10)$$

We can then obtain a corresponding representation of $SU_q(2)$ satisfying the algebra (S3) setting

$$K_s = q^{S^z_s} , \quad S^\pm_{s,q} |m\rangle = \sqrt{(s+1 \pm m)_q (s \mp m)_q} |m \pm 1\rangle \quad (S11)$$

where the subscript $q$ in $S^\pm_{s,q}$ has been added to distinguish them from the undeformed case $S^\pm_s$. Note that with respect to Eq. (S3), we replaced the abstract subscript $n$ with $s$, indicating the spin value. Here the notation

$$(x)_q = \frac{q^x - q^{-x}}{q - q^{-1}} \quad (S12)$$

has been introduced for the $q$-deformed integers.

These exhaust the possible irreducible representations of $SU_q(2)$ for generic values of $q$. However, when $q$ is a root of unity, there are some peculiarity. As explained in the text, here we focus on the case of principal roots of unity, i.e. $q = e^{\frac{2\pi i}{s}}$. One can verify that the representations in (S11) are no more irreducible when $s > \hat{s} = \frac{L-1}{2}$. Moreover for the maximal one, i.e. $s = \hat{s}$, there is a class of representations, parameterized by a complex parameter $\alpha$, which have the form

$$K_{\hat{s},\alpha} |m\rangle \equiv q^{m+\alpha} |m\rangle , \quad m = -\hat{s}, \ldots, \hat{s} , \quad (S13a)$$

$$S^+_{\hat{s},\alpha} |m\rangle = -(m - \hat{s} + 2\alpha)_q |m + 1\rangle , \quad (S13b)$$

$$S^-_{\hat{s},\alpha} |m\rangle = (m + \hat{s})_q |m - 1\rangle . \quad (S13c)$$

Note that for $\alpha \to 0$, this simply reduces to Eq. (S11) for $s = \hat{s}$.

These representations can be used to build the corresponding transfer matrices as explained in (S7) and (S9). We use the notation $T_{2\hat{s}}(\lambda) = \text{tr}_{2\hat{s}}(T_{2\hat{s}}(\lambda))$ for the operator associated with the representation of spin $s = 1/2, 1, \ldots, \hat{s} = (L-1)/2$. We use instead $T_{2\hat{s},\alpha}(\lambda)$ for the representation defined in (S13), as a function of the parameter $\alpha$. In general we will use $T_{2\hat{s}}(\lambda)$ to label all of them collectively.
B. EIGENVALUE OF THE CONSERVED CHARGES ON SINGLE-PARTICLE EIGENSTATES

Since all the transfer matrices commute \( T_{2s}(\lambda) \) among themselves for any pair of \( \lambda, \mu \) and \( s, s' \), it is possible to diagonalize all of them simultaneously. A special role is played by the transfer matrix \( T_1(\lambda) \) associated to the 2-dimensional fundamental representation. Once written as a \( 2 \times 2 \) matrix in the auxiliary space, it can be expressed in terms of four operators on the quantum space

\[
T_1(\lambda) = \begin{pmatrix} A(\lambda) & B(\lambda) \\ C(\lambda) & D(\lambda) \end{pmatrix}.
\] (S14)

so that \( T_1(\lambda) = A(\lambda) + D(\lambda) \). Then, the relation in Eq. (S8) provides the commutation relations between the entries which constitute the Yang-Baxter algebra. In particular, one can show that the simultaneous eigenstate of \( T_{2s}(\lambda) \) can be obtained by the multiple actions of \( B(\lambda) \) on the reference state \( |0\rangle = \uparrow \ldots \uparrow \) of all spin up

\[
|\lambda\rangle = B(\mu_1) \ldots B(\mu_M) |0\rangle
\] (S15)

provided that the Bethe-Ansatz equations (Eq. (3) in the main text) are satisfied for the rapidities \( \mu_i \). The eigenvalue of \( T_{2s}(\lambda) \) on the state \( |\lambda\rangle \) will be a symmetric function of the rapidities \( \mu_1, \ldots, \mu_M \). While for \( s = 1/2 \), the eigenvalue can be derived directly derived from the commutation relations deduced from Eq. (S8), for higher spin, the procedure is more involved. From the explicit expression of the transfer matrix in Eq. (S7) and the \( L \)-matrix in Eq. (S2), it is easy to obtain the eigenvalue on the reference state

\[
T_{2s}(\lambda) |0\rangle = \sum_{m=-s}^{s} (2 \sinh(\lambda + m\gamma))^L |0\rangle = \sum_{m=-s}^{s} f(\lambda + m\gamma) |0\rangle, \quad s = 1/2, \ldots, \hat{s}
\] (S16a)

\[
T_{2s,\alpha}(\lambda) |0\rangle = \sum_{m=-s}^{s} f(\lambda + i(m - \alpha)\gamma) |0\rangle
\] (S16b)

where the function \( f(z) = (2 \sinh(z))^L \). When many \( B(\lambda) \) operators act on the reference state In [26], this problem was solved for the standard representations by using the fact that higher transfer matrices \( T_{2s}(\lambda) \) with \( s > 1/2 \), can be obtained from the lowest one by tensor product on the auxiliary space. This procedure goes under the name of “fusion” and leads to a functional relation between the transfer matrices with different spin values

\[
T_{2s}(\lambda + \frac{i\gamma}{2})T_{2s}(\lambda - \frac{i\gamma}{2}) = f(\lambda + i(s + \frac{1}{2})\gamma)f(\lambda + i(s + \frac{1}{2})\gamma) + T_{2s-1}(\lambda)T_{2s+1}(\lambda), \quad s = 1/2, \ldots, \hat{s}.
\] (S17)

An explicit solution for this functional equation can be found as [26]

\[
T_{2s}(\lambda) = Q(\lambda + i(s + \frac{1}{2})\gamma)Q(\lambda - i(s + \frac{1}{2})\gamma) \sum_{m=-s}^{s} \frac{f(\lambda + m\gamma)}{Q(\lambda + i(m + \frac{1}{2})\gamma)Q(\lambda + i(m - \frac{1}{2})\gamma)}, \quad s = 1/2, \ldots, \hat{s}.
\] (S18)

where \( Q(\lambda) \) is the Baxter-Q operator, with eigenvalues:

\[
Q(\lambda) |\lambda\rangle = Q(\lambda; \{\mu_1, \ldots, \mu_M\}) |\lambda\rangle = \prod_{j=1}^{M} \sinh(\lambda - \mu_j) |\lambda\rangle.
\] (S19)

Combining Eqs. (S18) and (S19), one obtains the full spectrum \( T_{2s}(\lambda; \{\mu_1, \ldots, \mu_M\}) \) associated to standard representations \( s = 1/2, \ldots, \hat{s} \) and an arbitrary eigenstate. However, a similar approach does not seem to be immediately applicable for the \( \alpha \)-dependent maximal representation, as it cannot be obtained by fusing lower-spin ones.

Here, we follow a different approach. Driven by the simple generalization when \( \alpha \neq 0 \) in Eqs. (S16), we assume that \( T_{2s,\alpha}(\lambda) \) will still be expressed in terms of the \( Q \)-operator, with a structure similar to Eq. (S18). In order to find it explicitly, we turn to the simplest possible case: a state composed by a single rapidity \( \mu \). The eigenvalue on this kind of states can be obtained as the ratio

\[
T_{2s}(\lambda; \{\mu\}) = \langle \uparrow \ldots \uparrow | T_{2s}(\lambda) B(\mu) |0\rangle / \langle \uparrow \ldots \uparrow | B(\mu) |0\rangle.
\] (S20)

This equation is better represented graphically in Fig. S4, where the \( L \)-matrices are represented as boxes and the contractions over the indexes is indicated as edges joining two boxes. This representation suggests an efficient way
to compute it by performing preliminarily the contractions along each physical spin site (e.g. for the numerator the contraction $L_{1,2,1/2}(\mu)$ with $L_{s,1/2}(\lambda)$ along the vertical direction); then one is left with a matrix product in the auxiliary space (horizontal direction in Fig. S4), whose trace can be easily computed by diagonalizing the resulting operator. Using that $\mu$ satisfies the Bethe-Ansatz equations (which for a single particle reduce to a quantization condition), one obtains finally

$$T_{2s}(\lambda; \{\mu\}) = \sinh(\lambda + i(\hat{s} - \alpha + \frac{1}{2})\gamma) \sinh(\lambda - i(\hat{s} - \alpha + \frac{1}{2})\gamma) \sum_{m=-\hat{s}}^{\hat{s}} \frac{f(\lambda + i(m + \alpha)\gamma)}{\sinh(\lambda + i(m + \alpha + \frac{1}{2})\gamma) \sinh(\lambda + i(m + \alpha - \frac{1}{2})\gamma)}. \quad (S21)$$

As for a single rapidity $\mu$, the eigenvalue of the operator $Q(\lambda)$ are simply given by $\sinh(\lambda - \mu)$, it is natural to assume that Eq. (S21) generalizes to an arbitrary number of rapidities promoting $\sinh \to Q$, so that one arrives at the operator identity:

$$T_{2\hat{s},\alpha}(\lambda) = Q(\lambda + i(\hat{s} - \alpha + \frac{1}{2})\gamma)Q(\lambda - i(\hat{s} - \alpha + \frac{1}{2})\gamma) \sum_{m=-\hat{s}}^{\hat{s}} \frac{f(\lambda + i(m + \alpha)\gamma)}{Q(\lambda + i(m + \alpha + \frac{1}{2})\gamma)Q(\lambda + i(m + \alpha - \frac{1}{2})\gamma)}. \quad (S22)$$

We tested the correctness of this Ansatz numerically for two-rapidities states, but it remains a conjecture in the general case.

### C. Relation Between Conserved Quantities and Root Densities

The expression (S18) and its generalization (S22) can be used to obtain the eigenvalues of the trace of the transfer matrix in the thermodynamic limit $L \to \infty$. Indeed, because of the factor $f(\lambda + m\gamma)$ which involves an $L$-th power, for a given value of $\lambda$, only one sector (i.e. a single value of $m$) will exponentially dominate the sum. In particular, taking $\lambda$ in the neighborhood of the shift-point, i.e.: $\lambda \to \lambda + \gamma/2$, the sum will be dominated by the maximal $m = s$. We therefore define

$$X_{2s}(\lambda) = \frac{1}{2\pi i L} \frac{d}{d\lambda} \ln \left[ \frac{T_{2s}(\lambda + i\gamma/2)}{f(\lambda + i\gamma(s + \frac{1}{2}))} \right], \quad s = \frac{1}{2}, \ldots, \hat{s} \quad (S23)$$

$$X'(\lambda) = \frac{1}{2\pi i L} \frac{d}{d\alpha} \frac{d}{d\lambda} \ln \left[ \frac{T_{2\hat{s},\alpha}(\lambda + i\gamma/2)}{f(\lambda + i\gamma(s + \alpha + \frac{1}{2}))} \right]_{\alpha=0}. \quad (S24)$$

The operators $X_{2s}(\lambda)$ and $X'(\lambda)$ are generating functions for a complete set of local ($s = 1/2$) and quasi-local conserved charges in the XXZ spin chain. In particular, the logarithmic derivative with respect to $\lambda$ ensures that...
their eigenvalues in each eigenstate will be additive in the rapidities, leading to extensive expectation value for any operator in the Taylor expansion around \( \lambda = 0 \) (see Eq. (10) in the main text). The factor \( f(\ldots) \) in the denominator is chosen so that all of them have a vanishing eigenvalue on the reference state \( |0\rangle \), i.e., absence of any rapidity. In the thermodynamic limit, the rapidities in any eigenstate are arranged according to the string hypothesis [22]:

\[
\lambda_{n,v}^{a} = \lambda^{a} + \frac{r_{y}}{2}(n + 1 - 2a) + \frac{\pi(1 - v)}{4} + \delta_{p,a}^{a}, \quad a = 1, \ldots, n .
\]  

(S25)

In this expression, the index \( a \) labels the rapidities belonging to the same string, with the same real part \( \lambda^{a} \); the \( \delta_{p,a}^{a} \) is the deviation from the string hypothesis, which becomes exponentially small in the system size \( L \). String types are identified by the pair \( (n, v) \), with \( n \) the size of the string (number of rapidities) and \( v = \pm 1 \) is the string parity. In the gapless regime, whenever \( \Delta = \cos \gamma \), with \( \gamma \) is a rational multiple of \( \pi \) there exist only a finite number of string types. In particular, for the simplest case considered here, \( \gamma = \pi/\ell \), one has exactly \( \ell \) different types, \( j = 1, \ldots, \ell \) with \( n_{j} \) and \( v_{j} \)

\[
n_{j} = \begin{cases} j, & j = 1, \ldots, \ell - 1, \\ 1, & j = \ell \end{cases}, \quad v_{j} = \begin{cases} 1, & j = 1, \ldots, \ell - 1, \\ -1, & j = \ell \end{cases}
\]  

(S26)

In the thermodynamic limit the string momenta \( \lambda_{a}^{(j)} \) become dense on the real line and for each string type \( j \) we can introduce a density distribution \( L_{\rho}(\lambda_{a}^{(j)}) = (\lambda_{a+1}^{(j)} - \lambda_{a}^{(j)})^{-1} \) of them as a set of occupied (particles) and unoccupied (holes) root distributions \( \{\rho_{j}\}_{N_{j}=1}^{N} \cup \{\rho_{j}\}_{j=1}^{N} \), one for each string type. The two set of distributions are related by the thermodynamic version of the Bethe equations reading as

\[
v_{j}\rho_{j}(\lambda) = \alpha_{j}(\lambda) - \sum_{k=1}^{N_{j}} d\mu_{T_{jk}(\lambda - \mu)}(\mu) \quad j = 1, \ldots, N_{a} ,
\]  

(S27)

where \( \rho_{j}(\lambda) \equiv \rho_{j}(\lambda) + \rho_{j}^{a}(\lambda) \). Therefore only one of two sets, either the density of particles or of the holes, is sufficient to completely characterize the state in the thermodynamic limit. The sum over \( k \) runs over all the possible \( N_{a} \) types of particles with different parities \( \{v_{j}\}_{j=1}^{N_{a}} \) and lengths \( \{n_{j}\}_{j=1}^{N_{a}} \). We can then write the scattering kernels in (S27) as

\[
a_{n,v}(\lambda) = \frac{v}{\pi \cosh(2\lambda) - v \cos(\gamma n)} ,
\]  

(S28)

\[
T_{jk}(\lambda) = (1 - \delta_{n_{j}n_{k}})a_{|n_{j} - n_{k}|,v_{j}v_{k}}(\lambda) + 2a_{|n_{j} - n_{k}| + 2,v_{j}v_{k}}(\lambda) + \ldots + 2a_{n_{j} + n_{k} - 2,v_{j}v_{k}}(\lambda) + a_{n_{j} + n_{k},v_{j}v_{k}}(\lambda) .
\]  

(S29)

The conserved quantities can be evaluated on a thermodynamic state by considering the thermodynamic limit of their eigenvalue. These read as

\[
X_{2s}(x) | \rho \rangle = \sum_{j=1}^{\ell} \int_{-\infty}^{\infty} d\lambda \; q_{j}^{(s)}(x - \lambda) | \rho \rangle \quad , \quad s = \frac{1}{2}, 1, \ldots, \hat{s}
\]  

(S30a)

\[
X'(x) | \rho \rangle = \sum_{j=1}^{\ell} \int_{-\infty}^{\infty} d\lambda \; q_{j}'(x - \lambda) | \rho \rangle
\]  

(S30b)

In these expressions, the eigenvalues \( q_{j}^{(s)}(x) \) (and \( q_{j}'(x - \lambda) \)) corresponding to each \((n_{j}, v_{j})\)-string are obtained from those of a single rapidity, by summing over the whole string:

\[
q_{1}^{(s)}(x) = \frac{1}{2\pi} \frac{d}{dx} \ln \left[ \frac{\sinh(x - \frac{r_{y}}{2})}{\sinh(x + \frac{r_{y}}{2})} \right] = a_{2s,1}(x) \quad , \quad s = \frac{1}{2}, 1, \ldots, \hat{s} ,
\]  

(S31)

\[
q_{j}^{(s)}(x) = \sum_{a=1}^{n_{j}} q_{i}^{(s)} \left( x + \frac{r_{y}(n_{j} - 1 - 2a)}{2} + \frac{r_{y}(v_{j} - 1)}{4} \right) = \sum_{a=1}^{\min(n_{j},2s)} a_{2s - n_{j} - 1 + 2a,v_{j}}(x) .
\]  

(S32)

In a similar way, we can treat \( q_{j}'(x) \) and obtain

\[
q'(x) = \frac{d}{d\alpha}_{a_{2(\hat{s} - \alpha),1}(x)} \bigg|_{\alpha = 0} , \quad q_j'(x) = \frac{d}{d\alpha} \sum_{a=1}^{n_{j}} a_{2(\hat{s} - \alpha) - n_{j} - 1 + 2a,v_{j}}(x) \bigg|_{\alpha = 0}
\]  

(S33)
where we introduced the function

$$ a_{n,v}(\lambda) = \frac{v}{\pi} \frac{\sin(\gamma n)}{\cosh(2\lambda) - v \cos(\gamma n)} $$  \hfill (S34) 

Eqs. (S30) relates the expectation value of the full set of conserved quantities with the distributions of rapidities \( \{ \rho_1, \ldots, \rho_{\ell} \} \). The \( \ell - 1 \) generating functions of even charges \( X_{2s}(x) \), with \( s = \frac{1}{2}, 1, \ldots, \hat{s} \) can be used to fix \( \rho_1, \ldots, \rho_{\ell-2} \) and the difference \( \rho_{\ell} - \rho_{\ell-1} \). This is a consequence of the following relations between the eigenvalues

$$ q_j^{(s)}(x) = -q_{j-1}^{(s)}(x), \quad s = \frac{1}{2}, 1, \ldots, \hat{s} $$  \hfill (S35) 

which can explicitly checked in Eq. (S32). To simplify the notation we rearrange the \( \ell \) functions \( \rho_j, j = 1, \ldots, \ell \) as

$$ \tilde{\rho}_j(\lambda) = \begin{cases} 
\rho_j(\lambda) & j = 1, \ldots, \ell - 2 \\
\rho_{\ell-1}(\lambda) - \rho_{\ell}(\lambda) & j = \ell - 1 \\
\rho_{\ell}(\lambda) & j = \ell 
\end{cases} $$  \hfill (S36) 

and we can rewrite Eq. (S30a) as

$$ X_{2s}(x) |\rho\rangle = \sum_{j=1}^{\ell-1} \int_{-\infty}^{\infty} d\lambda \ q_j^{(s)}(x - \lambda) \tilde{\rho}_j(\lambda) |\rho\rangle, \quad s = \frac{1}{2}, 1, \ldots, \hat{s}. $$  \hfill (S37) 

We can now invert these relations. In order to do so, we observe that Eqs. (S30) and (S37) have the form of convolutions and becomes therefore multiplicative when going in Fourier transform. Defining for any function \( g(x) \), the Fourier transform \( \hat{g}(p) \) as

$$ \hat{g}(p) = \int_{-\infty}^{\infty} dx \ g(x) e^{ipx} $$  \hfill (S38) 

we have for the function \( a_{n,v}(\lambda) \) the expression

$$ \hat{a}_{n,v}(p) = \frac{\sinh\left( \frac{\pi x p}{2} - \frac{\pi x n}{2} \right)}{\sinh\left( \frac{\pi p}{2} \right)}, \quad k_v = \begin{cases} 
2 \left\lfloor \frac{\pi v}{\pi} \right\rfloor & v = 1 \\
2 \left\lfloor \frac{\pi v}{\pi} + \frac{1}{2} \right\rfloor & v = -1
\end{cases} $$  \hfill (S39) 

which holds for any \( n \in \mathbb{R} \) and \( v = \pm 1 \). Taking the Fourier transform of \( q_j^{(s)}(x) \) in Eq. (S32), we have

$$ \hat{q}_j^{(s)}(p) = \frac{\cosh\left( \frac{\pi x p}{2} - \frac{\pi x j - 2s - 1}{2} \right)}{2 \sinh\left( \frac{\pi p}{2} \right) \sinh\left( \frac{\pi p}{2} \right)} - \cosh\left( \frac{\pi x p}{2} (2s+1) \right) $$  \hfill (S40) 

Instead, in order to invert the relation (S30b) involving \( X'(x) \), we need the Fourier transform of \( q_j'(x) \)

$$ q_j'(p) = \frac{d}{dx} \sum_{\alpha=1}^{n_j} \hat{a}_{\ell-n_j+2(a-1)+2\alpha, v_1}(p) \bigg|_{\alpha=0} = \begin{cases} 
-\frac{\pi p \sinh\left( \frac{\pi p}{2} \right)}{\ell \sinh\left( \frac{\pi p}{2} \right) \tanh\left( \frac{\pi p}{2} \right)} & j = 1, \ldots, \ell - 1 \\
-\frac{\pi p \cosh\left( \frac{\pi (\ell-1)p}{2} \right)}{\ell \sinh\left( \frac{\pi p}{2} \right)} & j = \ell
\end{cases} $$  \hfill (S41) 

Finally, using the recursion relation

$$ \hat{q}_j^{(s+1/2)}(p) + \hat{q}_j^{(s-1/2)}(p) - 2 \cosh\left( \frac{\pi p}{2} \right) \hat{q}_j^{(s)}(p) = -\sinh\left( \frac{\pi p}{2} \right) \delta_{j,2s} $$  \hfill (S42) 

we arrive at the final set of equations for the \( \tilde{\rho}_j(p) \):

$$ \tilde{\rho}_j(p) - \delta_{j,\ell-1} \tilde{\rho}_\ell(p) = 2 \cosh\left( \frac{\pi p}{2\ell} \right) \hat{X}_j(p) - \hat{X}_{j+1}(p) - \hat{X}_{j-1}(p), \quad j = 1, \ldots, \ell - 1 $$  \hfill (S43) 

$$ \tilde{\rho}_\ell(p) = -\cosh\left( \frac{\pi p}{2\ell} \right) \hat{X}_{\ell-1}(p) - \frac{\ell}{\pi p} \sinh\left( \frac{\pi p}{2} \right) \hat{X}'(p) $$  \hfill (S44) 

These equations allow to fix a representative eigenstate \( |\rho\rangle \) in terms of the expectation values of all the local and quasi-local charges, generated by \( X_{2s}(x) \) and \( X'(x) \). In this way, as \( X_{2s}(x) \) and \( X'(x) \) remain constant throughout the quantum dynamics, they can be computed on the initial state and used to obtain the microcanonical GGE described by the corresponding \( |\rho\rangle \). This gives the complete characterization of any GGE state in the XXZ chain.
1. Spin-flip invariant GGE

Here we show that if the initial state $|\Phi\rangle$ has a definite parity under spin-flip, i.e.
\[ S |\Phi\rangle = \pm |\Phi\rangle , \]  
where the spin-flip operator $S = \left( \prod_{i=-\frac{L}{2}}^{\frac{L}{2}} s_i^z \right)$ as already introduced in the main text, then the constraint (S44) reduces to $\rho_{\ell} = \rho_{\ell-1}^h$. Under the spin-flip, the transfer matrix satisfies
\[ ST_{2s,0}(\lambda)S = T_{2s,0}(\lambda) . \]

From the definition of $X'(x)$ in (S24), we then deduce that
\[ X'(x) + SX'(x)S = \frac{1}{2\pi i L} \frac{d}{dx} \ln \left[ \frac{f(x + i\gamma(s - \alpha + \frac{1}{2}))}{f(x + i\gamma(s + \alpha + \frac{1}{2}))} \right] \]  
Then using Eq. (S45), we have for the expectation value of (S24) on $|\Phi\rangle$
\[ X_s'(x; \Phi) = \frac{1}{2\ell(\cosh x)^2} \Phi \text{ eigenstate of } S \]  
Now we recall that from the BA equations (equation (3) in the main text) one has
\[ \rho_j^h(x) = a_j(x) - X_j+1(x) - X_j-1(x) , \quad j = 1, \ldots, \ell - 1 \]  
and taking the difference with (S43) leads simply to
\[ \rho_j^h(x) + \delta_{j,\ell-1} \rho_{\ell}(x) = a_j(x) - X_j(x + \frac{i\gamma}{2}) - X_j(x - \frac{i\gamma}{2}) , \quad j = 1, \ldots, \ell - 1 , \]
Note here the additional term for $j = \ell - 1$, which simply comes from $\rho_{\ell-1}^h = \rho_{\ell-1} - \rho_{\ell}$. For $j = \ell - 1$, going to Fourier transform we have
\[ \hat{\rho}_{\ell-1}^h(p) + \hat{\rho}_\ell(p) = \hat{a}_{\ell-1}(p) - 2 \cosh(\frac{\pi p}{2\ell}) \hat{X}_j(p) \]  
Now, using (S48), and (S39) (we recall that $a_j = a_{n_j,\nu_j}$ and $n_{\ell-1} = \ell - 1$ and $\nu_{\ell-1} = 1$), we see that for any $|\Phi\rangle$ with definite spin-flip parity, we have:
\[ \hat{a}_{\ell-1}(p) = \frac{\sinh\left(\frac{\pi p}{2}\right)}{\sinh\left(\frac{\pi p}{2}\right)} = -\frac{2\ell}{\pi p} \sinh\left(\frac{\pi p}{2\ell}\right) \hat{X}'(p; \Phi) . \]  
Finally, inserting (S52) in (S51) and comparing the resulting expression with (S44), we see that for any state with definite parity we have
\[ \rho_{\ell-1}^h + \rho_{\ell} = 2\rho_{\ell} \quad \Rightarrow \quad \rho_{\ell} = \rho_{\ell-1}^h \]  

D. EVALUATION OF THE CONSERVED CHARGES ON A GENERIC PRODUCT STATE

It is easy to show that the decomposition of the transfer matrix $T_{2s}$ into a product of $L$ operators as in (S7) allows to compute the generating functions (S24) on a generic product state
\[ |\Phi\rangle = \bigotimes_i |\Phi_{i,i+p}\rangle \]  
where $|\Phi_{i,i+p}\rangle$ is a spin state for the spins between the position $i$ and $i+p$ with $p$ generic integer. Following [11] we introduce the $(2s + 1) \times (2s + 1)$ matrix of operators acting on the space $\mathbb{C}^2$ of the spin in position $i$
\[ L_{2s,i}(x, \mu) = \frac{L_{2s,a_i}(x + \mu)}{\sinh(x + i\gamma(s + \frac{1}{2}))\sinh(x + \mu - i\gamma(s + \frac{1}{2}))} . \]
where the product of the two Lax operators (defined in (S2)) is taken with respect to the auxiliary space indices. With this we define the following expectation value on one single constituent $|\Phi_{i,i+p}\rangle$ of the product state
\[ T_{2s}^{\Phi}(x,\mu) = \langle \Phi_{i,i+p} | L_{2s,i}(x,\mu) \cdots L_{2s,i+p}(x,\mu) | \Phi_{i,i+p} \rangle, \tag{S56} \]
which is still a matrix for dimension $(2s+1) \times (2s+1)$. Finally we define the generating function as traces over the auxiliary space indices
\[ X_{2s}(x) = \frac{1}{2\pi ip} \frac{\text{Tr}(\text{Adj}(T_{2s}(x,0) - 1) \mathbb{D}_{2s}^{\Phi}(x,0))}{\text{Tr}(\text{Adj}(T_{2s}(x,0) - 1))}, \tag{S57} \]
where $\mathbb{D}_{2s}^{\Phi}(\mu,x) = \partial_{\mu} T_{2s}^{\Phi}(x,\mu)|_{\mu=0}$ and the matrix coadjoint is defined as $\text{Adj}(A) \equiv \det(A)A^{-1}$. The same can be done for the generating function $X'(x)$ simply by introducing the also the derivative respect to $\alpha$. We define
\[ L_{2s,\alpha,\alpha}(x,\mu) = \frac{L_{(2s,\alpha,a_{1}}(x) L_{(2s,\alpha, a_{1})}^{\gamma}(x + \mu)}{\sinh(x + i\gamma(\delta + \alpha + \frac{1}{2})) \sinh(x + \mu - i\gamma(\delta - \alpha + \frac{1}{2}))}. \tag{S58} \]
such that
\[ T_{2s,\alpha}(x,\mu) = \langle \Phi_{i,i+p} | L_{2s,\alpha,\alpha}(x,\mu) \cdots L_{2s,\alpha,i+p}(x,\mu) | \Phi_{i,i+p} \rangle, \tag{S59} \]
With these elements we can then define the generating function $X'(x)$
\[ X'(x) = \frac{1}{2\pi ip} \frac{d}{d\alpha} \left[ \frac{\text{Tr}(\text{Adj}(T_{2s,\alpha}(x,0) - 1) \mathbb{D}_{2s,\alpha}^{\Phi}(x,0))}{\text{Tr}(\text{Adj}(T_{2s,\alpha}(x,0) - 1))} \right]_{\alpha=0}, \tag{S60} \]
with $\mathbb{D}_{2s,\alpha}^{\Phi}(\mu,x) = \partial_{\mu} T_{2s,\alpha}^{\Phi}(x,\mu)|_{\mu=0}$.

Note that the same method can be applied to evaluate the generating functions $X_{2s}$ and $X'$ when $|\Phi\rangle$ is a more generic matrix product state.