Role of impact ionization in the thermalization of photoexcited Mott insulators

Philipp Werner,[1,*] Karsten Held,[2] and Martin Eckstein[3]

[1]*Department of Physics, University of Fribourg, 1700 Fribourg, Switzerland*
[2]*Institut for Solid State Physics, Vienna University of Technology, 1040 Vienna, Austria*
[3]*Max Planck Research Department for Structural Dynamics, University of Hamburg-CFEL, Hamburg, Germany*

We study the influence of the pulse energy and fluence on the thermalization of photodoped Mott insulators. If the Mott gap is smaller than the width of the Hubbard bands, the kinetic energy of individual carriers can be large enough to produce additional doublon-hole pairs via a process analogous to impact ionization. The thermalization dynamics, which involves an adjustment of the doublon and hole densities, thus changes as a function of the energy of the photo-doped carriers and exhibits two time scales: a fast relaxation related to the impact ionization of high-energy carriers and a slower time scale associated with higher-order scattering processes. The slow dynamics depends more strongly on the gap size and the photodoping concentration.

PACS number(s): 71.10.Fd

I. INTRODUCTION

The photodoping of a Mott insulator provides a relatively simple way to induce and study a nonequilibrium phase transition. If a laser pulse with a frequency higher than the Mott gap is applied, doublon-hole pairs are produced, and these mobile carriers lead to a metallic response of the photodoped Mott insulator [1–3]. The changes in the optical conductivity associated with this metallization have been studied experimentally using time-resolved spectroscopy. In the pioneering work by Iwai and collaborators on a Ni-chain compound [1], a Drude peak in the conductivity was measured promptly after the photodoping pulse, and the metallic state was found to last for a few picoseconds. Alternatively, photoemission spectroscopy can be employed as a probe of the metallized Mott insulator, as shown for $1T - TaS_2$ [4,5].

One can distinguish two mechanisms which play a role in the relaxation of photodoped carriers: On the one hand, electron-electron scattering can lead to a thermalization of the electronic subsystem at a hot "electron temperature," and on the other hand, carriers can dissipate their initially high kinetic energy through scattering with "external" degrees of freedom such as spins or phonons. A large body of theoretical work on photodoped Mott insulators has focused on the latter relaxation processes involving scattering with spins in an antiferromagnetic background [6–9] or coupling to phonons [10–12]. In this paper we assume that electron-electron scattering is the fast mechanism, so that we can study the thermalization of isolated electrons and neglect the aforementioned energy loss processes, which affect the dynamics only on longer times; limits of this assumption are discussed in more detail below. Generally, this assumption is valid if the electron-phonon coupling strength is weak and if there are no spin correlations, e.g., due to a high temperature or large fluence.

In metals, rapid thermalization of the electronic system is typically observed and underlies the assumption that a quasiequilibrium picture or two-temperature model [13] can be used for describing the dynamics already at very short times

after an excitation. In an insulator, the thermalization and relaxation involve an adjustment in the number of electron-hole pairs, which can be a slow process in the presence of a large gap. It was found that in a purely electronic model (a paramagnetic Hubbard model with on-site repulsion U), the thermalization time depends exponentially on the gap size and that even the relaxation of the distribution of photodoped carriers within the Hubbard bands can be extremely slow [14]. The explanation for this is relatively simple. If the energy U which is needed for the production of a single doublon-hole pair is substantially higher than the typical kinetic energy of a single doublon or hole, complicated multiparticle scattering processes are needed for thermalization—hence the exponential scaling with U [15]. For a similar reason, the doublon-hole recombination via emission of magnons [16] or phonons [17] becomes slow when the gap is large.

If the kinetic energy of the charge carrier (doublon or hole) is higher than the size of the Mott gap, it is energetically allowed to create an *additional* doublon-hole pair via two-particle scattering. We call such processes "impact ionization," in analogy to similar processes in semiconductors [18] and atoms [19]. Since a high excess kinetic energy is needed, one may anticipate a strong dependence of the relaxation dynamics on the pulse energy (for a fixed interaction or gap size). In particular, we may encounter a situation where photodoped doublons inserted at the upper edge of the upper Hubbard band trigger a rapid increase in the number of charge carriers through impact ionization, while for doublons inserted at the lower band edge, the kinetic energy is not sufficient for impact ionization, so that the doublon-hole production depends on rare multiparticle scattering events.

Since impact ionization processes have the potential to rapidly enhance the number of mobile carriers in the photoinduced metal, an understanding of this physics is crucial for possible applications of photoinduced metal-insulator transitions in ultrafast switches or for the efficient operation of photovoltaic devices. For example, impact ionization allows us to create multiple doublon-hole pairs per photon and hence to overcome the Schockley-Queisser limit for the efficiency of solar cells [20,21].

In this paper we focus on a paramagnetic Mott-Hubbard insulator with a relatively small gap and study in more

detail the electronic thermalization processes, with the goal of disentangling the fast impact ionization channel from doublon-doublon scattering and slower multiparticle scattering processes. As we will show, impact ionization can be identified experimentally by characteristic signatures in the time-resolved photoemission spectrum, in particular, its energy and fluence dependence: (i) Impact ionization can increase the number of doublons on the 10-fs time scale and, thus, result in a rapid spectral weight increase of up to a factor of 3 above the Fermi level. (ii) While the photoemission spectral weight at high energies decreases as a function of time, it increases more than proportionally at energies which are lower by at least the size of the gap. (iii) The frequency of the pump pulse needs to be higher than twice the gap; below this threshold, impact ionization is not possible. (iv) Since impact ionization involves only a single doublon or hole this process does not depend strongly on the density of photodoped carriers (or the fluence), while doublon-doublon and higher order scattering processes will become more frequent if the density of photoinduced carriers increases.

II. MODEL AND METHOD

We investigate and quantify the effect of impact ionization by considering a Hubbard model,

$$H = \sum_{ij,\sigma} v_{ij} c_{i\sigma}^\dagger c_{j\sigma} + U \sum_i \left(n_{i\uparrow} - \frac{1}{2}\right)\left(n_{i\downarrow} - \frac{1}{2}\right), \quad (1)$$

with on-site interaction U comparable to the bandwidth. The operators $c_{i\sigma}$ create an electron at site i with spin σ, and the hopping amplitude is v_{ij}. The model is solved on an infinite-dimensional hypercubic lattice using nonequilibrium dynamical mean-field theory (DMFT) [22,23] with a strong-coupling perturbative impurity solver (noncrossing approximation NCA) [24]. This lattice has a Gaussian density of states, $\rho(\epsilon) = 1/(\sqrt{\pi} W) \exp(-\epsilon^2/W^2)$, and we use the width W as our unit of energy. To simulate the photo-doping pulse, we apply a few-cycle electric field pulse of the form

$$E(t) = E_0 e^{-(t-t_p)^2/\sigma^2} \sin(\Omega(t - t_p)), \quad (2)$$

with $t_p = 6$ and $\sigma^2 = 6$ in the body-diagonal of the lattice. We use a gauge without a scalar potential, so that the field is given by the time derivative of the vector potential A, $E(t) = -\partial_t A(t)$. Using the Peierls substitution, the field then enters the Hamiltonian via a time-dependent shift of the dispersion, $\epsilon_k \to \epsilon_{k-A(t)}$, where ϵ_k is the Fourier transform of the hopping matrix. For details on our implementation of the nonequilibrium DMFT equations, the treatment of the electric field, and the noncrossing approximation impurity solver, we refer to Refs. [14] and [23]. The double occupation at time t, $d(t)$, is a local observable which can be obtained directly from the solution of the effective impurity model. Since we use a strong-coupling impurity solver, $d(t) = \mathcal{G}_{|\uparrow\downarrow\rangle}^<(t,t)$ is simply the occupation of the pseudoparticle state corresponding to doubly occupied sites [24]. Because d is nonzero already in the initial Mott state due to virtual charge fluctuations, the number of photoexcited doublons at a later time t is approximately given by the difference $d(t) - d(0)$. This number can also be

obtained spectroscopically, by integrating the photoemission spectrum over the upper Hubbard band (see below).

The relaxation dynamics of photodoped carriers depends crucially on whether or not the Mott insulator is antiferromagnetically ordered [7,8,25]. In the present study, we restrict the calculations to paramagnetic photodoped Mott insulators at elevated temperatures. The temperature considered in this paper, i.e., $T = 1/\beta = 1/5$, is well above the highest DMFT Néel temperature, $T_N = 1/7$ [26]. Since DMFT overestimates T_N, we can be certain that also paramagnons are not important. This is also supported by recent nonequilibrium cluster DMFT calculations [8], which, for the two-dimensional Hubbard model, showed that short-ranged spin correlations have an important effect on the relaxation dynamics only below $T = (1/5) 4t_*$ [27]. (For higher-dimensional lattices, the effect of short-range correlations in the paramagnetic phase can be expected to be even smaller.)

III. RESULTS

A. Pulse-frequency dependence

For the purpose of orientation we first plot the equilibrium spectra for different values of U and inverse temperature $\beta = 5$ (Fig. 1). The gap opens at $U \approx 2.5$ and then grows approximately linearly with U. In the insulating phase, the shape of the Hubbard bands is almost independent of U, and they have a width of about 3. Since an impact ionization process involves the scattering of a doublon at the upper edge of the upper band to the lower edge and a simultaneous doublon-hole excitation, it already becomes clear from the equilibrium spectra that we can only expect these processes to be relevant for interactions $U \lesssim 4$. For larger U values, the energy associated with scattering between states within the band is not enough to excite electrons across the gap. In the following we thus focus on the interaction range $2.5 \leqslant U \leqslant 4$.

For $U = 2.5$ and 3.5, the time evolution of the photodoped doublon density D after pulses with frequencies in the range $1.5\pi/2 \leqslant \Omega \leqslant 3.5\pi/2$ is plotted in the top panels in Fig. 2. Here and in the following, we examine the change
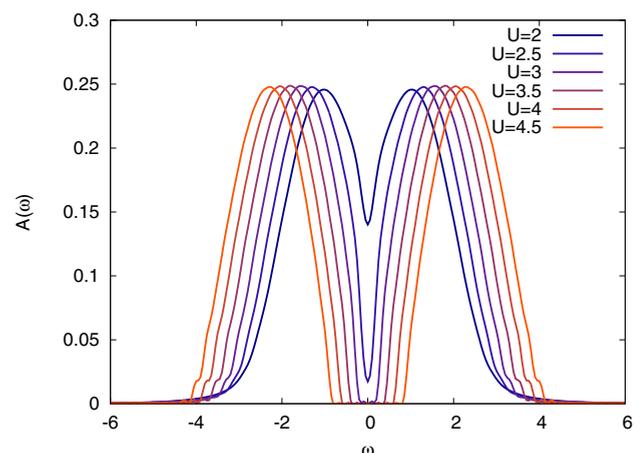

FIG. 1. (Color online) Equilibrium spectral functions of the Hubbard model for $\beta = 5$ and indicated values of U as calculated by DMFT (with NCA impurity solver).
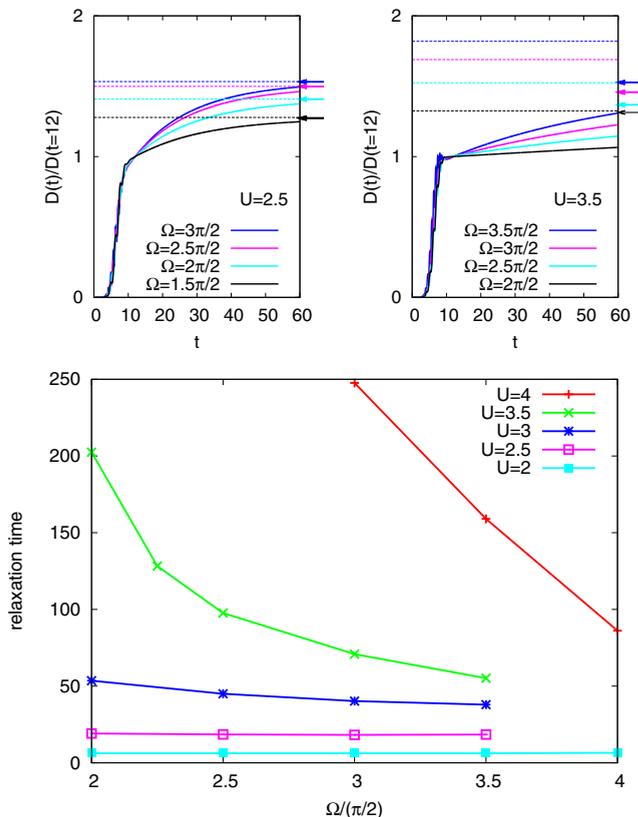
FIG. 2. (Color online) Relaxation after pulse excitations at different frequencies at an initial inverse temperature $\beta = 5$. The amplitude of the pulses is adjusted such that the number of photodoped doublons at $t = 12$ (shortly after the pulse) is 0.01. Top: Time evolution of the normalized doublon density and expected thermal values (horizontal lines) for $U = 2.5$ and 3.5. Bottom: Relaxation times (c) obtained by fitting $D(t)/D(t = 12)$ to the function $a + b \exp(-t/c)$ in the range $t \in [30,60]$. The extrapolated long time values (a) from this fit are indicated by arrows in the upper panels.

of double occupation with time, $D(t) = d(t) - d(0)$. For a better comparison between different band gaps and pulse energies, the amplitude E_0 of each pulse has been adjusted such that at $t = 12$, shortly after the pulse, the density of photodoped doublons is $D(t = 12) = 0.01$, and we normalize the curves by this initial density. We see that during the thermalization process, the number of doublons increases, i.e., excess kinetic energy of the photodoped carriers is transformed into interaction energy. The thermal reference value can be calculated by measuring the energy $E_j = \int dt\, j(t) \cdot E(t)$ injected into the system by the pulse. Here, $j = \sum_k n_{k\sigma} v_k$ is the current, with $n_{k\sigma}(t) = -i G^<_{k\sigma}(t,t)$ and $v_k(t) = \partial_k \epsilon_{k-A(t)}$. By comparing the total energy after the pulse to that of an equilibrium system, we can compute the temperature $1/\beta_{\text{eff}}$ and double occupancy which the system will reach, assuming thermalization, in the long-time limit. The thermal values of the double occupancy are indicated by the dashed horizontal lines in Fig. 2.

If we fit the doublon curves in the range $30 \leqslant t \leqslant 60$ to a single exponential $a + b \exp(-t/c)$ we obtain the relaxation times c plotted in the bottom panel in Fig. 2 and the long-time values a indicated by the arrows in the top panels. For $U \gtrsim 3$,

this extrapolated thermalized double occupation a is smaller than its actual thermal value (dashed line in Fig. 2). Hence, we can conclude that at least two relaxation mechanisms are at work. We also note that the relaxation times c are much faster than previously observed electronic thermalizations of doublons [14,15] and strongly pulse-energy dependent: with increasing pulse frequency, the initial growth of the doublon population becomes faster (see upper right and lower panels in Fig. 2). All of this suggests that the fast doublon production is due to impact ionization, which requires that the excess kinetic energy of the photodoped carriers is higher than the gap. Once all carriers with a high kinetic energy have decayed, this contribution disappears and the long-time thermalization dynamics is controlled by slower multiparticle scattering processes. In previous studies, this was the only relaxation mechanism, since the pulse energy was too low (relative to the gap) for impact ionization.

In order to show direct evidence for impact ionization, we plot in Fig. 3 time-resolved photoemission spectra [28],

$$I(\omega,t) = -i \int dt_1 dt_2 S(t_1) S(t_2) e^{i\omega(t_1 - t_2)} G^<(t + t_1, t + t_2),$$

$$(3)$$

for a Gaussian probe pulse envelope $S(t) = \exp(-t^2/2\delta^2) \Theta(1.5\delta - |t|)$, which we cut off for $|t| > 1.5\delta$. Choosing $\delta = 12$ and a pump-pulse lasting up to $t = 12$, this means that for $t > 30$, there is no longer any overlap between the pump and the probe pulse. A pulse width of δ allows us to measure the relaxation of the system with an energy resolution of $\sim 1/\delta$.

Let us first focus on the left panels in Fig. 3, which show results for interaction $U = 3$, pulse amplitude 2, and pump pulse frequency $\Omega = 3.5\pi/2$. This pump pulse inserts the doublons at the upper edge of the upper Hubbard band. As the time increases, the spectral weight near the upper band edge decreases, while the weight near the lower band edge starts to increase. [Up to a constant offset, the total weight in the upper Hubbard band reproduces $d(t) - d(0)$ to a high accuracy.] Since the energy difference between the upper and the lower band edge is larger than the gap size in this example, impact ionization processes can be expected to play a role in the initial relaxation dynamics.

That these processes are indeed, to a large extent, responsible for the doublon production follows from the lower panel, which plots the difference between the photoemission spectrum at time t and the measurement at time $t = 24$. The spectral weight decreases with increasing time for $\omega \gtrsim 1.9$, with the fastest decrease measured at energy $\omega_{\text{loss}} = 2.55$ (indicated by an arrow in Fig. 3). From the lower band edge up to $\omega \lesssim 1.9$, we see an increase in spectral weight, with the fastest doublon production at energy $\omega_{\text{gain}} = 0.85$. Impact ionization can now be identified by analyzing the number of doublons produced per decay of a high-energy doublon. Let us consider the process of a high-energy doublon creating a doublon-hole pair (doublon$_{\text{high}} \rightarrow$ doublon$_{\text{low}} +$ doublon$_{\text{low}} +$ hole$_{\text{low}}$) and its symmetric counterpart (hole$_{\text{high}} \rightarrow$ hole$_{\text{low}} +$ hole$_{\text{low}} +$ doublon$_{\text{low}}$). The net effect is the production of *three* low-energy doublons per decay of a high-energy doublon. Therefore, if impact ionization were the only relevant process on the time scale in Fig. 3, we would expect that the increase in the
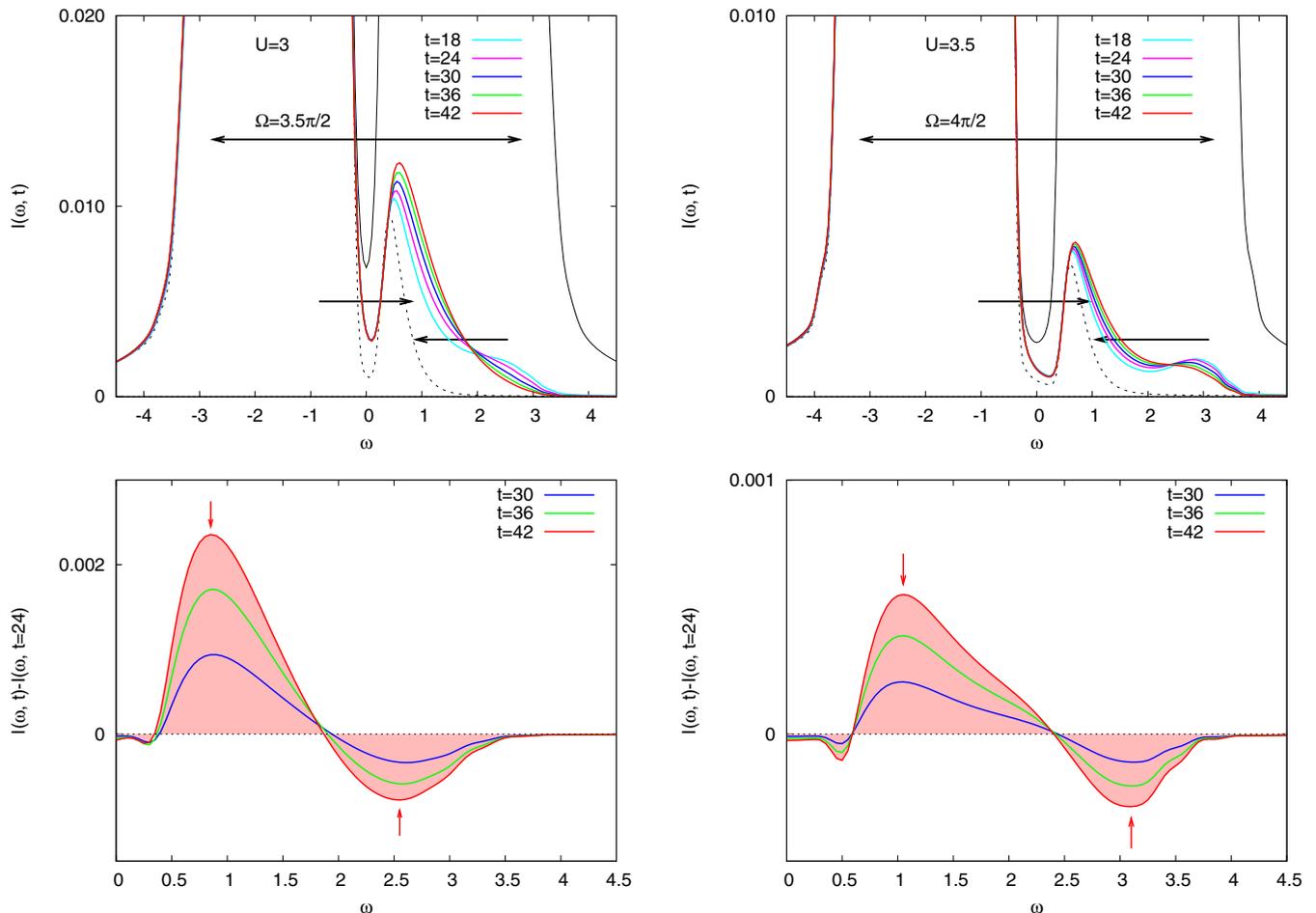
FIG. 3. (Color online) Time-resolved photoemission spectra for pulse amplitude 2 and initial inverse temperature $\beta = 5$. Left: Results for $U = 3$, $\Omega = 3.5\pi/2$; initial photodoping concentration $D(t = 12) = 0.0056$. Right: Results for $U = 3.5$, $\Omega = 4\pi/2$; initial photodoping concentration $D(t = 12) = 0.0021$. Colored curves in the upper panels show the nonequilibrium photoemission spectrum $I(\omega,t)$ for indicated values of t, while dashed black curves plot that of the initial equilibrium state. Solid arrows sketch the energy transfers associated with an impact ionization process: left-pointing arrow—kinetic energy loss of a high-energy doublon; right-pointing arrow—excitation of an electron across the gap. Bottom: Time-dependent change in the photoemission spectrum. The (red) arrows indicate the energies $\omega_{\text{gain}} = 0.85$ and $\omega_{\text{loss}} = 2.55$ ($U = 3$; left) and $\omega_{\text{gain}} = 1.05$ and $\omega_{\text{loss}} = 3.1$ ($U = 3.5$; right). Shaded (red) areas correspond to the increase (decrease) in low-energy (high-energy) doublons from $t = 24$ to $t = 42$. For $U = 3$, the increase in the number of low-energy doublons is about 2.7 times as large as the decrease in the number of high-energy doublons, while for $U = 3.5$, the corresponding factor is about 2.3.

number of low-energy doublons would be three times larger than the decrease in the number of high-energy doublons. Computing the integrals over the positive and negative parts of the curves displayed in the bottom panel [shaded (red) areas], we find a ratio of 2.7. This indicates that besides the impact ionization processes, there are also doublon-conserving scattering processes which contribute to the redistribution of spectral weight within the Hubbard band. We discuss some key differences between these two relaxation channels, and how they affect the time-resolved photoemission spectra, in the section on the fluence dependence (Sec. III C).

A second observation is that the ratio in the positions of the maxima of the gain and loss peak is approximately $\omega_{\text{gain}}/\omega_{\text{loss}} = 3$. A tempting interpretation would be to say that in an impact ionization process the shift of occupied spectral weight within the upper Hubbard band from ω_{loss} to ω_{gain} is associated with a transfer of occupation between the Hubbard bands from $-\omega_{\text{gain}}$ to ω_{gain} (scattering process

indicated by arrows in the upper panel). However, in contrast to a band insulator the first moment of the occupied density of states in Mott insulators does not equal the total energy, so that the aforementioned redistribution of occupied weight would not be energy conserving. Doublon-hole excitations lead to a reconstruction of the density of states and hence a redistribution of weight over a larger ω region. Here we do not analyze this effect in detail but, instead, focus on the evolution of the spectral weight averaged over large energy regions (high- and low-energy doublons).

The right-hand panels in Fig. 3 show analogous results for $U = 3.5$, pulse amplitude 2, and pulse frequency $\Omega = 4\pi/2$. While the absorption is lower in this case, the parameters are still compatible with impact ionization. Indeed, as shown in the lower panel, the change in the spectral function is fastest near the energies $\omega_{\text{gain}} = 1.05$ and $\omega_{\text{loss}} = 3.1$, which satisfy $\omega_{\text{gain}}/\omega_{\text{loss}} \approx 3$. The low-energy hump is, however, broader, and the ratio between the shaded (red) areas is only 2.3, which

suggests a larger role of doublon-doublon and doublon-hole scattering processes in this case.

B. Two-step thermalization

At least in cases such as the setup discussed above, where the high-energy and low-energy carriers can be relatively clearly separated, one can try to reproduce the time evolution of the doublon population with a simple model that describes the decay of the high-energy doublons via impact ionization with a relaxation time γ and the higher order scattering processes with a different associated thermalization time τ. We denote the slow processes with the subscript "therm" and the fast ones with "imp" and split the total doublon number D into a high-energy and a low-energy population, D_1 and D_2, respectively. After thermalization, we assume that only low-energy doublons are present and denote their number D_{th}. The time evolution is then given by the equations $\frac{dD_1}{dt} = (\frac{dD_1}{dt})_{\text{imp}}$ and $\frac{dD_2}{dt} = (\frac{dD_2}{dt})_{\text{therm}} + (\frac{dD_2}{dt})_{\text{imp}}$, where we assume the simple rate equations

$$\left(\frac{dD_1}{dt}\right)_{\text{imp}} = -\frac{1}{\gamma}D_1, \qquad (4)$$

$$\left(\frac{dD_2}{dt}\right)_{\text{imp}} = -3\left(\frac{dD_1}{dt}\right)_{\text{imp}}, \qquad (5)$$

$$\left(\frac{d}{dt}D_2\right)_{\text{therm}} = \frac{1}{\tau}\left(D_{\text{th}} - D_2\right). \qquad (6)$$

The factor of 3 in Eq. (5) accounts for the production of three low-energy doublons per decay of a high-energy doublon (hole) in an impact ionization process, as explained above. The equations governing the time evolution of the two components thus read $\frac{dD_1}{dt} = -\frac{1}{\gamma}D_1, \frac{dD_2}{dt} = \frac{1}{\tau}(D_{\text{th}} - D_2) + \frac{3}{\gamma}D_1$, and the solution for the total doublon population for times $t > t_s$ becomes

$$D_{\text{th}} - D(t) = \frac{2\tau + \gamma}{\tau - \gamma}D_1(t_s)e^{-(t-t_s)/\gamma}$$
$$+ \left(D_{\text{th}} - D(t_s) - \frac{2\tau + \gamma}{\tau - \gamma}D_1(t_s)\right)e^{-(t-t_s)/\tau}. \qquad (7)$$

Here, t_s is some time after the pulse (we choose $t_s = 15$ in the following analysis), D_{th} and $D(t_s)$ are known, and $D_1(t_s)$, γ, and τ must be obtained by fitting.

For $U = 2.5$ the relaxation is well described by a single exponential. This follows already from the data in the top left panel in Fig. 2, which show that the extrapolated long-time values from an exponential fit in the range $t \in [30,60]$ correctly predict the thermal doublon density. This is, however, a special case, since the gap is just opening at $U = 2.5$. In this situation, additional doublons can be easily generated and the relaxation to the expected thermal value is fast.

For $U = 3$ and 3.5, a single-exponential model is not appropriate anymore, but fitting of the data with the double-exponential decay, Eq. (7), works rather well. We summarize the results of this analysis in Table I. One finds fast relaxation times of $\gamma \sim 15$ and slow relaxation times of $\tau \sim 60$ for $U = 3$, and fast (slow) relaxation times of approximately 40–50 (250–350) for $U = 3.5$. The (relative) initial excited

TABLE I. Relaxation times and initial excited populations $D_1(t_s)$ extracted from fits to model (7) in the range $t \in [15,60]$ ($t_s = 15$). The doping concentration after the pulse is $D(t_s) = 0.010$ in all cases.

U	Ω	$D_{\text{th}} - D(t_s)$	$\frac{D_1(t_s)}{D(t_s)}$	γ	τ
2.5	$\frac{3\pi}{2}$	0.00448	0.0088	7.20	18.8
2.5	$\frac{2.5\pi}{2}$	0.00421	0.0067	7.75	19.0
2.5	$\frac{2\pi}{2}$	0.00348	0.0044	9.35	19.6
3	$\frac{3.5\pi}{2}$	0.00684	0.046	13.4	60.3
3	$\frac{3\pi}{2}$	0.00674	0.040	15.0	61.4
3	$\frac{2.5\pi}{2}$	0.00573	0.026	16.5	64.9
3.5	$\frac{3.5\pi}{2}$	0.00789	0.15	44.0	376
3.5	$\frac{3\pi}{2}$	0.00669	0.083	48.4	257
(4	$\frac{4\pi}{2}$	0.00820	0.19	86.9	5990)

population decreases as the pulse frequency is lowered, in rough agreement with the time-resolved spectra. At the lowest pulse frequencies considered, the separation between high-energy and low-energy populations becomes blurred and our model fit becomes less meaningful. For $U = 4$, all relaxation times become rather long, and it is difficult to obtain reliable fits. We find $\gamma \approx 90$ and $\tau \approx 6000$ (with a large uncertainty).

While one should probably not consider more than the first digit of the relaxation times and initial high-energy populations in Table I, our model does provide a consistent description of the doublon relaxation, and the results demonstrate that impact ionization processes play a significant role in the interaction range $3 \leqslant U \leqslant 4$. In particular, they lead to a two-step thermalization with a fast initial doublon production and an associated transfer of spectral weight from the upper to the lower band edge, followed by a much slower thermalization of the relaxed distribution. The slow time scale τ grows rapidly with increasing U, which is consistent with a previous analysis based on single-exponential fits [14]. However, also the relaxation time γ associated with the impact ionization increases with U, which indicates that these processes become less likely as the energy cost of producing a doublon-hole pair increases. Note that the excess kinetic energy of the doublon has to be higher than the Mott gap, which increases with U. On the other hand, the kinetic energy of the photodoped carriers is essentially bounded by the noninteracting bandwidth, which is independent of U.

An instructive way to illustrate the two-step relaxation is to plot the doublon production rate $(d/dt)D(t)$ as a function of the deviation of the doublon density from the thermal value, $D_{\text{th}} - D(t)$. In this case, our model predicts a crossover from a small linear slope (corresponding to the slow long-time thermalization process) to a steeper slope (corresponding to the impact ionization processes). Indeed, for $U \gtrsim 3$, the data sets for different pulse energies fall roughly onto a single curve which describes such a crossover (upper panel of Fig. 4). For the data sets corresponding to the highest pulse frequencies, we plot the fits to model (7) with dashed lines. These fits also roughly reproduce the relaxation for the other pulse frequencies, which shows that the model provides a consistent description of the thermalization process.

In the lower panel in Fig. 4, we show the time evolution for $U = 3.5$, $\Omega = \frac{3\pi}{2}$, as predicted by the model (parameters
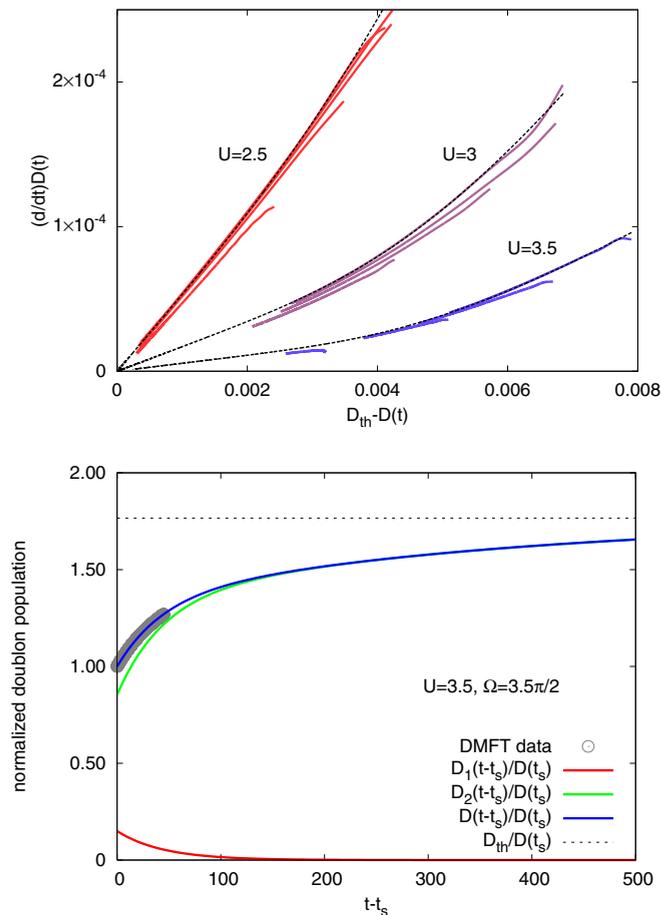
FIG. 4. (Color online) Top: Doublon production rate plotted as a function of $D_{th} - D(t)$ for different pulse energies and indicated values of U. Fits of model (7) to the curves corresponding to the highest pulse energy are shown by dashed lines. Bottom: Time evolution of the normalized doublon population as predicted from the fit to model (7) for $U = 3.5$, $\Omega = 3.5\frac{\pi}{2}$ ($t_s = 15$).

from Table I). One can clearly see the two-step relaxation to the thermal value (dashed line), with a rapid initial increase in the doublon density, linked to impact ionization, followed by a much slower thermalization. Even though the relative high-energy population is small (about 15% at $t_s = 15$), the impact ionization process contributes about half of the additional doublons needed for thermalization.

C. Fluence dependence

The impact ionization processes can be distinguished from the slower thermalization processes also by analyzing the dependence of the relaxation times on the photodoping concentration, or fluence. Since impact ionization involves only a single doublon or hole in the initial state, we expect a weak fluence dependence of the fast relaxation time γ. On the other hand, the higher-order scattering processes that increase the number of doublons involve several doublons and/or holes. Hence these processes should exhibit a stronger dependence on the photo-doping concentration, so that we expect an increase in the slow relaxation time τ as the pulse amplitude is decreased.

TABLE II. Relaxation times and initial populations extracted from fits to model (7) in the range $t \in [15,60]$ for $U = 3$, $\Omega = 3.5\pi/2$, $t_s = 15$, and the indicated pulse amplitudes.

Amplitude	$D(t_s)$	D_{th}	β_{th}	γ	τ
0.25	0.000108	0.000236	4.884	19.9	214
0.5	0.000429	0.000917	4.593	19.5	194
1	0.00167	0.00334	3.879	18.3	147
2	0.00593	0.0105	2.854	15.6	85.0
6	0.0165	0.0252	1.996	11.2	46.1

We analyze the fluence dependence of the relaxation for $U = 3$, $\Omega = 3.5\pi/2$, and pulse amplitudes ranging from 0.25 to 6. The doping concentrations in the thermalized state and at $t = 15$, shortly after the pulse, are given in Table II. For a low pulse amplitude, the number of photodoped carriers grows proportionally to the square of the pulse amplitude, as expected. The thermalization in this regime leads to more than a doubling of the mobile carriers. For pulse amplitudes $\gtrsim 2$, the number of carriers grows more slowly than the power of the field pulse, and also the relative increase in the doublon population associated with thermalization is smaller. To avoid complications due to strongly nonlinear absorption processes, we do not consider higher amplitudes.

The top panel in Fig. 5 shows the corresponding time evolution of the doublon concentration, normalized at $t = 15$. The results for amplitudes <2 all collapse onto a single curve. This shows that in the initial stage of the relaxation, the doublon-hole production becomes independent of the doping concentration—a result consistent with a time evolution which is dominated by impact ionization. To see that the slow time scale is indeed more strongly dependent on the fluence, we plot in the lower panel the difference from the thermal value, $D_{th} - D(t)$, on a logarithmic scale. To extract the two relaxation times, γ and τ, we performed fits of the DMFT data with model (7) and $t_s = 15$. The slow time scale τ increases from about 50 to about 200 as the pulse amplitude is lowered from 6 to 0.25, while the fast time scale increases from about 10 to about 20 (Table II). For pulse amplitudes lower than 2, i.e., in the low doping regime, the fast time scale becomes essentially independent of the doping concentration, while the slow time scale shows no sign of saturation and continues to increase with decreasing doping concentration. [The relative high-energy population, $D_1(t_s)/D(t_s)$, also increases.] The estimated value of $\gamma \approx 20 \ll \tau$ implies that the initial fast increase in the doublon population evident in the upper panel in Fig. 5 is due to impact ionization.

One also finds a fluence dependence in the time-resolved photoemission spectra. If the bandwidth, gap size, and pump-pulse energy are compatible with impact ionization, then these processes dominate the doublon production and redistribution of spectral weight if the density of carriers is low. As discussed in Sec. III A, a characteristic signature in the photoemission spectrum is an increase in spectral weight at an energy ω_{gain} near the lower band edge and a simultaneous decrease at an energy ω_{loss}, where about three low-energy doublons are produced per high-energy doublon. This behavior is clearly evident for pulses with a low amplitude in Fig. 6. The top
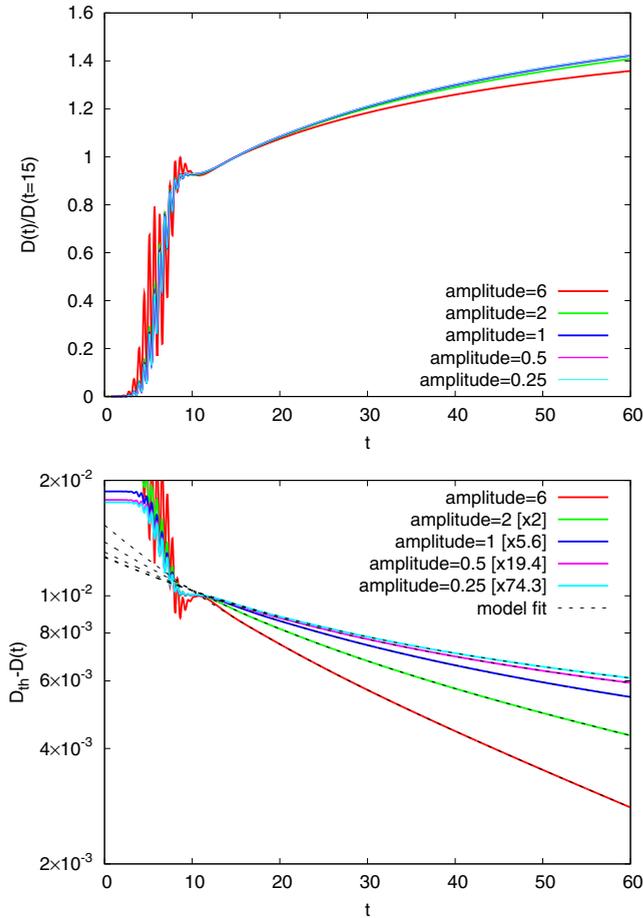
FIG. 5. (Color online) Time evolution of the doublon concentration for $U = 3$, $\Omega = 3.5\pi/2$, and different pulse amplitudes. Top: Normalized doublon population. Bottom: Relaxation of the doublon concentration to the thermal value. The curves for amplitudes <6 are multiplied by an arbitrary factor, to enable a better comparison of the long-time behavior. Dashed lines are fits to model (7) in the time interval $t \in [15, 60]$.
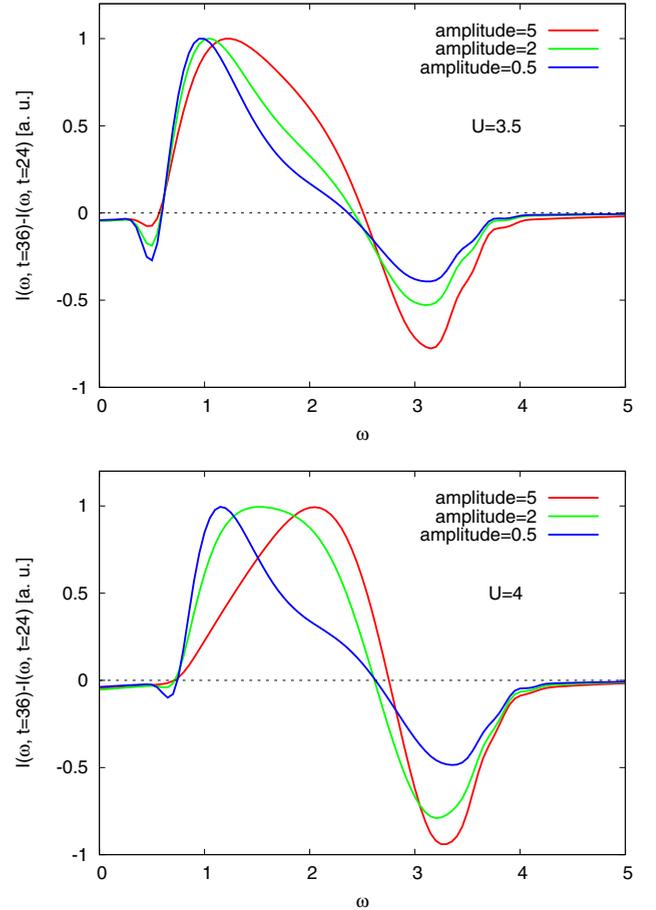

FIG. 6. (Color online) Difference between the time-resolved photoemission spectra measured at $t = 36$ and $t = 24$ for $\Omega = 4\pi/2$ and the indicated values of the pulse amplitude. Top: $U = 3.5$. Bottom: $U = 4$. For better comparison, curves have been normalized such that the maximum difference is 1. For $U = 3.5$, the doublon concentration at $t = 12$ is $D = 0.0065$, 0.0021, and 0.00015 for amplitude 5, 2, and 0.5, respectively. For $U = 4$, the corresponding numbers are $D = 0.014$, 0.0043, and 0.00030.

panel shows results for $U = 3.5$ and $\Omega = 4\pi/2$. For the blue curve [$D(t = 12) = 0.00015$], the area under the positive hump near the lower band edge is 2.5 times larger than the area under the negative hump at high energies. For the green curve [$D(t = 12) = 0.0021$] it is 2.3 time larger. This means that for each doublon which disappears at high energy, more than two doublons are created at low energy, and thus we conclude that most of the doublons appearing near ω_{gain} are produced by impact ionization. A similar result is also displayed in the lower panel ($U = 4$, $\Omega = 4\pi/2$), where for the lowest pulse amplitude [$D(t = 12) = 0.0003$] we see two well-defined peaks with $\omega_{\text{gain}}/\omega_{\text{loss}} \approx 3$.

With increasing fluence, the low-energy hump broadens, which means that doublons appear at energies in the middle of the band, which are no longer compatible with impact ionization. They are instead the result of scattering processes between high-energy and low-energy doublons (or holes) which conserve the number of carriers. The red curve in the upper panel in Fig. 6, which corresponds to $D(t = 12) = 0.0065$, and the green curve in the lower panel [$D(t = 12) = $

0.0043] show that these processes become relevant already at a doublon density of ∼0.5%. In these simulations, the area under the low-energy hump is about a factor of 2 larger than the area under the negative hump, which indicates that for each impact ionization process $\text{doublon}_{\text{high}} \rightarrow 3 \text{ doublons}_{\text{low}}$, we have approximately also one scattering between a high-energy and a low-energy doublon (or hole), $\text{doublon}_{\text{high}} + \text{doublon}_{\text{low}} \rightarrow 2 \text{ doublons}_{\text{intermediate}}$. Because the scattering probability of the latter process is proportional to the carrier concentration, it can dominate the redistribution of spectral weight at an even higher fluence (see the red curve in the bottom panel).

We note that deviations from the universal low-fluence evolution of the relative doublon concentration also appeared for doping concentrations higher than ∼0.5% (see Fig. 5 and Table II). Thus our analysis of the photoemission spectra supports the interpretation that the universal curve is entirely controlled by impact ionization, while the slower increase in the relative doublon population seen for higher fluences is the result of competing scattering processes which deplete the high-energy population.

IV. DISCUSSION

In this study, we considered the thermalization dynamics after a photodoping pulse in a purely electronic model without magnetic order. For a photoexcited Mott insulator, it is instructive to distinguish (i) relaxation processes within the Hubbard band (which keep the number of doublon-hole pairs fixed) and (ii) relaxation processes across the Mott gap (which change the number of doublon-hole pairs). The simplest relaxation process of type (i) is electron-electron scattering, which here is a doublon-doublon, hole-hole, or doublon-hole scattering. It keeps the number of doublons and holes fixed but transfers energy from one doublon (hole) to another. This process requires a second doublon (or hole), and hence the corresponding relaxation rate of a carrier will be proportional to the number of doublons (holes). Such processes redistribute the energy among the doublons (holes) but do not change the doublon number.

In order to change the number of doublons, type (ii) processes are needed. Again, the simplest process is electron-electron scattering, which in this case corresponds to impact ionization: a doublon (or hole) excites an electron across the Mott gap, creating an additional doublon-hole pair. From one doublon, we obtain two doublons and one hole. If we also consider the symmetric process for holes, impact ionization leads to a threefold increase in the number of doublons and holes. These processes do not involve other doublons (or holes) but the doublon kinetic energy must exceed the size of the Mott gap. If the doublon energy is not large enough, only less likely multiscattering events can thermalize the number of doublons.

We have discussed the characteristic signatures of impact ionization in situations where the pulse energy is high and the gap is small, as in Fig. 3: In this situation the photoexcited high-energy doublons create additional doublon-hole pairs so that the number of doublons almost triples, and a second peak develops in the photoemission spectrum at an energy corresponding to about one-third of the photoexcited high-energy peak. Our data analysis based on the model, Eqs. (4)–(6), assumes that the high-energy doublon population D_1 decays only via impact ionization and gives a good fit in these cases. We also find that impact ionization is fast for a Mott insulator. For a bandwidth W of the order of 1 eV, corresponding to a unit of time of 0.66 fs, the fast relaxation times in Table I are of the order of 5–70 fs.

We have also seen that processes of type (i) have the potential to prevent impact ionization by lowering the doublon energy before impact ionization occurs. Doublon-doublon scattering processes become more important when the number of doublons is large (Fig. 6). In this case the rate, Eqs. (4) and (5), should be extended to $(dD_1/dt)_{\text{imp+scat}} = -(1/\gamma)D_1 - (1/\eta)D_1 D_2$ and $(dD_2/dt)_{\text{imp+scat}} = +(3/\gamma)D_1 + (1/\eta)D_1 D_2$, where η is the relaxation time for scattering processes of type (i).

In a real material further relaxation processes not considered in our paper are possibly important, in particular, phonon and (para)magnon scattering. In many cases, Mott gaps are of the order of 1 eV so that phonons and magnons have a lower energy and hence can contribute only to type (i) processes. But as discussed above, this has the potential to prevent impact ionization. In a Mott insulator with strong electron-phonon

coupling, the cooling rate associated with electron-phonon scattering can be of the same order of magnitude as impact ionization [12]. For most systems electron-phonon relaxation occurs, however, only on the 0.1- to 1-ps time scale, which means that these processes are slower than the observed impact ionization in a Mott insulator. This is completely opposite to the behavior in semiconductors, where impact ionization has a much larger time scale than electron-phonon scattering [20,30]. For semiconductors, electron-phonon coupling hence prevents impact ionization altogether.

Even in a purely electronic system, additional relaxation processes may come into play. At low temperatures, in the magnetically ordered phase, spin-flip scattering provides a particularly efficient dissipation channel, which can lead to a fast redistribution of spectral weight within the Hubbard bands. Exact diagonalization-based studies of the motion of a single carrier in an antiferromagnetic background suggest that the excess kinetic energy of a photodoped carrier is transferred to the spin background within a few hopping times [9], and recent DMFT studies of photodoped antiferromagnetic Mott insulators revealed a very fast cooling of the photocarriers [7,25]. Also, in the vicinity of an antiferromagnetic phase, short-range spin correlations provide an efficient scattering mechanism [8], whereas in one dimension the energy transfer to the spin system seems to be inefficient [29].

In a model which takes into account the absorption of excess doublon kinetic energy by phonon or magnon scattering, we have a reduction in the high-energy population D_1 and a corresponding increase in the low-energy population D_2. However, this time these processes do not depend on the number of doublons, hence Eq. (4) has to be modified as $(dD_1/dt)_{\text{imp+ph/mag}} = (-1/\gamma - 1/\kappa)D_1$, and Eq. (5) as $(dD_2/dt)_{\text{imp+ph/mag}} = (3/\gamma + 1/\kappa)D_1$, where κ is the corresponding relaxation time. The efficient dissipation of kinetic energy and the associated rapid decrease in the high-energy population in an antiferromagnetic system is expected to have a significant effect on the thermalization dynamics in small-gap Mott insulators. It reduces the effectiveness of the impact ionization process and leads to a slower adjustment of the doublon population and, thus, a slower electronic thermalization. In the present work, however, we have chosen a high temperature at which spin correlations are reduced, so that impact ionization can be more clearly identified. Our results should also be relevant at higher fluences, independent of magnetic ordering, because in this case the photodoping leads to a rapid melting of antiferromagnetic correlations [7].

V. CONCLUSION AND OUTLOOK

The main finding of this study is that in situations where the gap size is smaller than the width of the Hubbard bands, the kinetic energy of the photodoped particles can be high enough that impact ionization processes play an important role in the initial relaxation. In fact, for the largest interactions considered ($U = 3.5$–4), the doublon-hole production on the computationally accessible time scales is almost entirely due to impact ionization processes. We have demonstrated this by analyzing the time-resolved photoemission spectrum and by extracting the impact ionization and thermalization time scales from fits to a model with two exponentials, which

was found to provide a rather good description of the time evolution of the doublon density. These time scales depend on the gap size, with the slow time scale (related to higher order scattering processes) growing much more rapidly with gap size than the fast one (related to impact ionization), while the pulse frequency mainly affects the relative population of high-energy carriers which can trigger impact ionizations. The two time scales also exhibit a different dependence on the pulse amplitude (or density of photodoped carriers): impact ionization processes are insensitive to the doping concentration in the low-doping regime, while the slow time scale grows rapidly with decreasing fluence. For higher photodoping, impact ionization can be masked and suppressed by doublon-doublon scattering.

Impact ionization may be relevant for Mott solar cell applications. In the case of conventional semiconductor solar cells, the Coulomb interaction is weak so that interaction scattering (impact ionization) can hardly excite an electron across the semiconducting gap, i.e., create an additional electron-hole pair. Impact ionization only occurs on time scales of 1–100 ps [30], which is much longer than the typical time scales of 0.1–1 ps for electron-phonon scattering. Hence, for a conventional semiconductor, almost all the excess kinetic energy of photoinduced carriers is transferred to lattice vibrations (heat). Consequently, for each photoexcited electron-hole pair, only the gap size is harvested as an electrical energy quantum, independent of the energy quantum of the photon. This severely restricts the efficiency of semiconductor solar cells to about 31%, known as the Schockley-Queisser limit [20]. To overcome this limit, solar cell quantum dots, e.g., based on PbSe nanocrystals, where larger Coulomb interactions and phonon-bottlenecks effects can enhance impact ionization have been proposed [31]. Also, for Mott insulators the

possibility of impact ionization has been discussed based on Fermi's Golden Rule calculations for the Hubbard model [21].

In our paper we have shown that impact ionization in a Mott insulator can occur on time scales of the order of 10 fs, i.e., much faster than typical electron-phonon relaxation times. Impact ionization processes are efficient only in Mott insulators with a small gap relative to the width of the Hubbard bands. This is quite difficult to realize for a one-band Hubbard model. However, in multiband Hubbard models or in charge-transfer insulators, the size of the gap can be much smaller than the width of the Hubbard and charge transfer bands, respectively. Whether or not impact ionization can contribute significantly to the power produced by Mott solar cells such as the recently proposed $LaVO_3$-based heterostructure [32] remains an open question. To address this issue one would have to consider a realistic setup and, also, study the diffusion of the photodoped carriers to the leads [33], the effect of the spin background [7], and the coupling to phonons [34]. In any case, our study has shown that impact ionization in Mott insulators can be fast and can contribute effectively to the production of carriers. Hence, Mott insulators have the potential to overcome the Schockley-Queisser limit, by harvesting more than the gap energy per photon. This class of materials can thus be expected to play an important role in the future development of highly efficient solar cells.

ACKNOWLEDGMENTS

[1] S. Iwai, M. Ono, A. Maeda, H. Matsuzaki, H. Kishida, H. Okamoto, and Y. Tokura, Phys. Rev. Lett. **91**, 057401 (2003).

[2] H. Okamoto, H. Matsuzaki, T. Wakabayashi, Y. Takahashi, and T. Hasegawa, Phys. Rev. Lett. **98**, 037401 (2007).

[3] H. Okamoto, T. Miyagoe, K. Kobayashi, H. Uemura, H. Nishioka, H. Matsuzaki, A. Sawa, and Y. Tokura, Phys. Rev. B **82**, 060513 (2010).

[4] L. Perfetti, P. A. Loukakos, M. Lisowski, U. Bovensiepen, H. Berger, S. Biermann, P. S. Cornaglia, A. Georges, and M. Wolf, Phys. Rev. Lett. **97**, 067402 (2006).

[5] J. C. Petersen, S. Kaiser, N. Dean, A. Simoncig, H. Y. Liu, A. L. Cavalieri, C. Cacho, I. C. E. Turcu, E. Springate, F. Frassetto, L. Poletto, S. S. Dhesi, H. Berger, and A. Cavalleri, Phys. Rev. Lett. **107**, 177402 (2011).

[6] J. Kogoj, Z. Lenarcic, D. Golez, M. Mierzejewski, P. Prelovsek, and J. Bonca, arXiv:1402.6104.

[7] M. Eckstein and P. Werner, Phys. Rev. Lett. **113**, 076405 (2014).

[8] M. Eckstein and P. Werner, arXiv:1410.3956.

[9] D. Golez, J. Bonca, M. Mierzejewski, and L. Vidmar, Phys. Rev. B **89**, 165118 (2014).

[10] D. Golez, J. Bonca, L. Vidmar, and S. A. Trugman, Phys. Rev. Lett. **109**, 236402 (2012).

[11] H. Matsueda, Sh. Sota, T. Tohyama, and S. Maekawa, J. Phys. Soc. Japan **81**, 013701 (2012).

[12] P. Werner and M. Eckstein, arXiv:1403.7376.

[13] Ph. B. Allen, Phys. Rev. Lett. **59**, 1460 (1987).

[14] M. Eckstein and P. Werner, Phys. Rev. B **84**, 035122 (2011).

[15] R. Sensarma, D. Pekker, E. Altman, E. Demler, N. Strohmaier, D. Greif, R. Jördens, L. Tarruell, H. Moritz, and T. Esslinger, Phys. Rev. B **82**, 224302 (2010).

[16] Z. Lenarcic and P. Prelovsek, Phys. Rev. Lett. **111**, 016401 (2013).

[17] M. Mitrano, G. Cotugno, S. R. Clark, R. Singla, S. Kaiser, J. Stähler, R. Beyer, M. Dressel, L. Baldassarre, D. Nicoletti, A. Perucchi, T. Hasegawa, H. Okamoto, D. Jaksch, and A. Cavalleri, Phys. Rev. Lett. **112**, 117801 (2014).

[18] W. Shockley, Czech. J. Phys. B **11**, 81 (1961); L. V. Keldysh, Sov. Phys. JETP **21**, 1135 (1965).

[19] T. D. Märk and G. H. Dunn, *Electron Impact Ionization* (Springer-Verlag, Wien, Germany, 1985).

[20] W. Shockley and H. J. Queisser, J. Appl. Phys. **32**, 510 (1961).

[21] E. Manousakis, Phys. Rev. B **82**, 125109 (2010).

[22] J. K. Freericks, V. M. Turkowski, and V. Zlatić, Phys. Rev. Lett. **97**, 266408 (2006).

[23] H. Aoki, N. Tsuji, M. Eckstein, M. Kollar, T. Oka, and P. Werner, Rev. Mod. Phys. **86**, 779 (2014).

[24] M. Eckstein and P. Werner, Phys. Rev. B **82**, 115115 (2010).

[25] P. Werner, N. Tsuji, and M. Eckstein, Phys. Rev. B **86**, 205101 (2012).

[26] M. Jarrell, Phys. Rev. Lett. **69**, 168 (1992).

[27] Here, $4t_*$ is half the bandwidth in two dimensions, and $t_* \equiv 1$ is the unit of energy in Ref. [8].

[28] J. K. Freericks, H. R. Krishnamurthy, and T. Pruschke, Phys. Rev. Lett. **102**, 136401 (2009).

[29] K. A. Al-Hassanieh, F. A. Reboredo, A. E. Feiguin, I. González, and E. Dagotto, Phys. Rev. Lett. **100**, 166403 (2008).

[30] M. Wolf, R. Brendel, J. H. Werner, and H. J. Queisser, J. Appl. Phys. **83**, 4213 (1998); A. P. Kirk and M. V. Fischetti, Phys. Rev. B **86**, 165206 (2012).

[31] A. J. Nozik, Annu. Rev. Phys. Chem. **52**, 193 (2001); Physica E **14**, 115 (2002); R. D. Schaller and V. I. Klimov, Phys. Rev. Lett. **92**, 186601 (2004); R. D. Schaller, V. M. Arganovich, and V. I. Klimov, Nat. Phys. **1**, 189 (2005).

[32] E. Assmann, P. Blaha, R. Laskowski, K. Held, S. Okamoto, and G. Sangiovanni, Phys. Rev. Lett. **110**, 078701 (2013).

[33] M. Eckstein and P. Werner, Phys. Rev. B **88**, 075135 (2013).

[34] P. Werner and M. Eckstein, Phys. Rev. B **88**, 165108 (2013).