Clarity and Positioning. While we were somewhat surprised about the comments in this regard (we dedicated 3-4 pages to just discussions and positioning, including nearly a page of related work), we now believe after careful considerations of the reviews that some of these discussions did not appear in the places where readers might expect them, e.g., closer to theorem statements. Given the page limits we had instead opted to discuss proof ideas close to theorem statements but should have reiterated and referenced previous discussions. We will move proof ideas to the appendix, and heed the various suggestions. In terms of related work, we strived to diligently overview the vast literature, although we missed some as pointed out by one reviewer (Reviewer 3); we however emphasize that the suggested references, while relevant, are somewhat peripheral to our main contributions (see below).

Significance and Contributions. Our main contributions are twofold:
(1) We derive the *first* general measures of discrepancy (transfer exponents) **that yield tight rates in transfer** for convergence of excess risk to zero. We argue that these new measures overcome significant limitations suffered by other discrepancies previously proposed in the literature, thereby describing many scenarios where transfer is possible but where this fact is not implied by any previously-defined notions of discrepancy ($d_{\mathcal{A}}$, L_p, KL-divergence, Kernel-mean discrepancy, Wasserstein). This is due e.g. to the fact that transfer is inherently an *asymmetric problem* (one distribution might have sufficient information on the other but not the other way around), so cannot be captured by metrics (L_1, $d_{\mathcal{A}}$, Wasserstein, Kernel-mean discrepancy), and does not require that distributions have the same support (KL-divergence, density ratios). The recent notion of [18] has a similar flavor, but is concerned with specific smoothness assumptions on the regression function $\mathbb{E}[Y|X]$, and does not have a way to take into account the structure of an abstract hypothesis class $\mathcal{H}$: for instance, [18] still requires distributions with the same support (as discussed in Related work section, and Example 1, lines 77-80 and 124-131). Indeed, our starting point for this work was to try to extend the insights of [18] to yield a general $\mathcal{H}$-dependent discrepancy measure (for the purpose of overcoming shortcomings in the well-known $d_{\mathcal{A}}$ and $d_{\mathcal{Y}}$ discrepancies, as we discuss in Example 2).

(2) We show for the first time that many questions of a practical interest can in principle be addressed in a near-optimal way, *without estimating the discrepancy between distributions*, but through clever use of unlabeled data. For instance, nearly all previous works on transfer assumed no labeled target data, and therefore could not address practical questions such as the optimal mix of source and target labeled data (of potentially different sampling costs); in this regard, we give the first bounds **in terms of any mix of n_P source and n_Q target labeled samples for VC classes** and show that these bounds are tight. What's surprising here is that the optimal rate behaves as $\min\{\mathcal{E}_Q(\hat{h}_P), \mathcal{E}_Q(\hat{h}_Q)\}$: i.e., is akin to ignoring one of the source or target data sets (but note that we still need to use both samples in order to decide which one to ignore in the end). This is then essential in assessing optimal sampling under mixed costs.

We *mitigate the usual worst-case nature of minimax analysis* by showing that our bounds are **tight for any given hypothesis class**, and, **tight in any noise regime** (Theorems 1 and 2). We are not aware of any previous such individual tightness result in the context of transfer.

What we left open. In this first step, we mainly focused on understanding the theoretical limits of transfer under mixed sampling costs, and outlined various theoretical procedures (non-implementable as they are based on ERM on 0-1 loss), which yield the insights that various practical questions can be addressed in principle with no prior distributional knowledge: i.e., from data-driven decisions alone. Given the simplicity of the algorithmic principles outlined, a natural next step is to investigate practical versions of these procedures based on convex surrogate losses, and extensions to address multiclass learning, regression, and other general prediction settings.

Other questions by reviewers. Space limits prevent answering all reviewer questions, but we address a few below.
Reviewers 1 and 2. The marginal transfer exponent γ indeed depends on unknown h^*_P. However we never have to estimate it, as we show that the ERM $\hat{h}_P$ on source data is a good surrogate for h^*_P. Hence, using *cheap* unlabeled data, we can identify good hypotheses h using the fact that $Q_X(h \neq \hat{h}_P) \approx Q_X(h \neq h^*_P)$ which upper-bounds the excess error $\mathcal{E}_Q(h)$ (this is all done rigorously in the paper). Unfortunately, using unlabeled data to drive choices, all we can adapt to is the marginal exponent which, unlike ρ, requires little knowledge of $Q_{Y|X}$. Note however that if all we are interested in is the basic transfer problem of using n_P and n_Q given samples (sections 4 and 5), then we can adapt optimally to even ρ without estimating it (Theorem 3), which is surprising as we know very little about $Q_{Y|X}$ when n_Q is small. Regarding the question about $\mathcal{P}$, any set defined a-priori is fine (not data-dependent).

Reviewer 3. We hope many of your concerns were addressed above. The references mentioned will be added. They use metric notions of discrepancy (see above), and in contrast to our work, largely leave open the question of tightness.

Reviewer 4. The results allow $P_{Y|X} = Q_{Y|X}$: i.e., we do not *require* them to be different in the upper bounds (and there exist classes where the lower bounds hold even with restriction to $P_{Y|X} = Q_{Y|X}$). As per Theorem 4, few target labels are required whenever γ is small and $\mathfrak{c}_P \ll \mathfrak{c}_Q$; we will add relevant comments to the paper. The results hold for *any* VC class, and are instantiated simply by plugging in the VC dimension of the class.
We thank you for pointing out the typo in line 249. "C" is meant to be $\tilde{C}$ and in fact should double at each iteration.