# Modeling large-scale structure

Martin White

(UC Berkeley, LBNL)

Benasque, August 2023

# Outline

- The golden age of cosmological surveys.
  - We are gathering vast amounts of data that can inform us about many interesting questions in physics and astrophysics ... given suitable models!
- Modeling the evolution of large-scale structure
  - Analytic approaches.
  - Numerical simulations (see O.Hahn's talk).
  - Better together!
- Concluding thoughts.

# Golden age of surveys

We are living in the "golden age" of cosmological surveys, with survey capabilities increasing exponentially ... (Moore's law)

- DESI has completed 2 (of 5) years of operation.
- PFS is commissioning, WEAVE will begin soon.
- ACT has completed its 'final' observations.
- Euclid is at L2!
- Simons Observatory is under construction (Adv SO approved)
- LSST will be coming online in the next few years
- SPHEREx and Roman will launch later in the decade
- CMB-S4 will follow in the next decade
- ... and others.

Each is powerful in its own right, together they will be amazing!

# Modeling challenge

We are gathering vast amounts of data that can inform us about many interesting questions in physics and astrophysics ... given suitable models!

Model requirements:

- complete, i.e. capable of modeling each of our tracers/probes of large-scale structure including their cross-correlations,
- consistent (among different observables),
- accurate (to percent level or better),
- well-controlled (uncertainty quantification),
- with high dynamic range in redshift and scale, to break degeneracies.

# Models of large-scale structure (LSS)

How do people model measurements of large-scale structure?

- There are two broad classes of approaches to modeling LSS: analytical and numerical.
- Analytic approaches based on perturbation theory (PT) – which have seen a renaissance in recent years and will become increasingly powerful with future surveys.
- Numerical approaches (simulations) – the workhorse.
- New ideas for combining the two: "best of both worlds".

# Perturbation theory (PT)

- Cosmology deals with relativistic gauge field theories, like many other sub-fields of physics.
- The equations of motion are both non-linear and non-local.
- PT developed starting in the 1960's, reached its "classical" form in the early 1990's (with important developments to this day).
- Standard techniques familiar from QM, condensed matter or particle theory
  - Effective field theory
  - Green's functions, diagrams, tree level, loops, ...
  - Regularization, renormalization, running, counter terms, IR resummation, ...

# Sort-of like QFT

We can make this look a lot like QFT (or stat.mech., fluid mech.!)

- Collect density, velocity, etc. into a vector $\varphi^a$.
- Can rewrite the EOM as 'propagation' and 'interaction'.
- Now rather than a Feynman path integral for operator expectation values, have ensemble averages of "initial" fields:

$$\left\langle \varphi^a \cdots \varphi^b \right\rangle = \int \mathcal{D}\phi_{\rm ic}\, \varphi^a[\phi_{\rm ic}] \cdots \varphi^b[\phi_{\rm ic}] \exp \underbrace{\left[ -\frac{1}{2} \phi^i_{\rm ic} \{ P^{-1}_{ij} \} \phi^j_{\rm ic} \right]}_{S_0[\phi_{\rm ic}]}$$

  that can be obtained by functional derivatives of (log of)

$$Z[J] = \int \mathcal{D}\phi_{\rm ic}\ \exp \left\{ S_0[\phi_{\rm ic}] + J_i \varphi^i[\phi_{\rm ic}] \right\}$$

- Many techniques carry across directly, though there are some technical differences.

# A tale of two expansions

Note: in "standard" perturbation theory we are dealing with two expansions – one for the dynamics and one for galaxy bias.

A full model requires both, but they are conceptually different.

# Analytic models: PT

- Perturbation theory provides clean predictions for
  - matter (lensing) and biased tracers (galaxies, QSOs, ...)
  - in real and redshift space.
  - pre- and post-reconstruction
- Robust to uncertainties in small-scale physics ("integrated out").
  - No additional assumptions about halos, galaxies, etc. needed beyond the (minimal) set of bias parameters dictated by fundamental symmetries.
- Consistent predictions of $P_\ell(k)$, $\xi_\ell(s)$, $C_\ell^{\kappa g}$, $C_\ell^{\kappa\kappa}$, $w(\theta)$, ... using the same parameters.

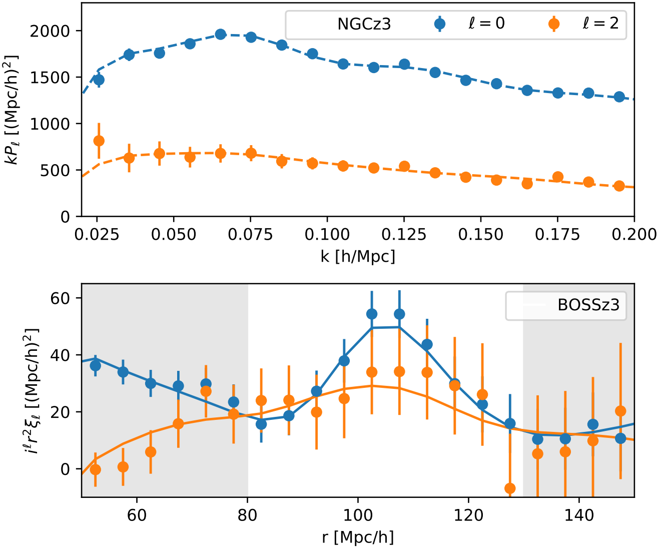
Widely used: most of the constraints on cosmological parameters from galaxy redshift surveys over the last decade have come from these kinds of analytic models!

# Velocileptors

- We have a public, Python package for these models.
- Being used in a number of surveys and data analyses now.
- Many ways to combine velocities and densities in power spectra: direct PT expansion, moment expansion, Gaussian streaming model, Fourier streaming model.
- Available in both LPT and EPT variants (allowing cross-checks!)
- Works in Fourier and configuration space.
- Fast and “easy to use”; works with several analysis packages.

http://github.com/sfschen/velocileptors

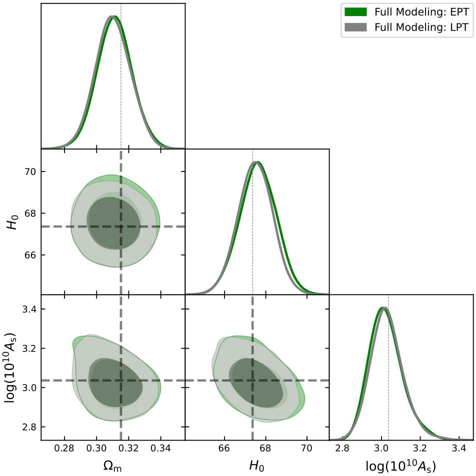
# Models fit current (BOSS) data well




Chen et al. (2022)

# DESI “full shape” modeling challenge

A comparison of Eulerian and Lagrangian models fit to mock $\Lambda$CDM data:




Maus+23

# DESI “full shape” modeling challenge

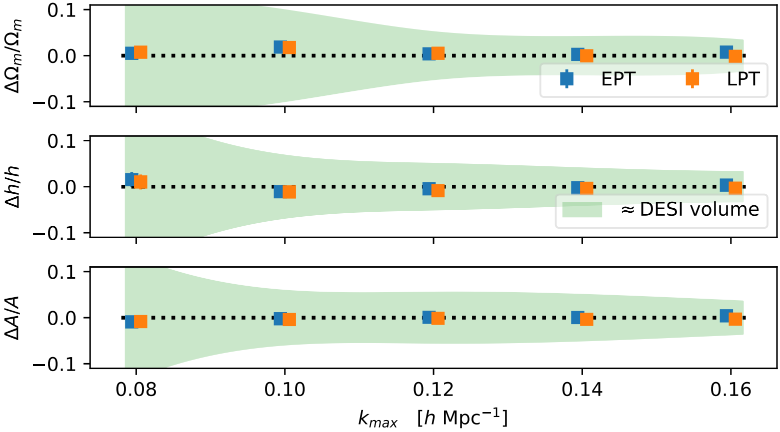
Within DESI we are running an additional test on the PT codes:

- Each model fits a particular mock dataset, with the cosmology fixed to ‘truth’.
- The best-fit model is then passed as ‘noiseless’ mock data to the other codes.
- The other codes fit these data to provide the cosmology contours.

The agreement between the different groups is very good so far!

# PT blind challenge

Inferring parameters from fits to mock survey data 100× larger than physically achievable volumes:




"PT blind challenge", Nishimichi+20

# 3D vs 2D

- For 3D analyses PT is likely more than adequate for the foreseable future.
- Have "plenty" of modes even at low $k$, and modeling to high $k$ is costly.
  - Need to model 'stochastic terms' (fingers of god), including satellites. Introduces many parameters, little cosmological information left (?).
  - Limited by redshift errors (for some tracers and some experiments).
- For 2D measurements have a different problem: $N_{\rm modes} \propto \ell_{\rm max}^2$.
- No problem with FoG, but always "mode starved".
- Want to push to as small a scale as can get cosmology from!
- PT is limited in reach – can never model small scales no matter what order you compute to!

# Numerical simulation

- The other approach is to simulate the formation of structure through Monte-Carlo integration of the Vlasov equation (N-body simulations).
- This is probably the dominant approach in the field right now, and the most familiar!
- The N-body problem is well defined:
  - Generate a (Gaussian) 'initial' density/potential field at some suitable starting redshift, $z_{\mathrm{init}}$.
  - Use low-order LPT to compute positions and velocities for a 'grid' of particles (so they're in the density field growing mode).
  - Repeat as long as necessary:
    - Use positions to compute accelerations.
    - Use accelerations to update velocities
    - Use velocities to update positions.

(See O. Hahn's excellent talk from last week!)

# Numerical simulation

- The addition of other collisionless species (e.g. neutrinos) is conceptually straightforward, if numerically challenging.
- Non-radiative hydrodynamics is a well-posed problem, though codes still disagree more than we'd like.
- Including cooling and feedback is the research frontier ... the best way to deal with the complexities of galaxy formation, is still an open research problem.
  - hard to scale to cosmological volumes.
  - hard to tune to match observations.
  - use of phenomenological subgrid models poses interesting questions of convergence and completeness.
  - Uncertainty quantification an open problem.
- While more expensive simulations are a "UV complete" model (of something) and can predict 'any' observable.

Simulation and analytic approaches are not in conflict – can we gain anything from combining them?

# Better together

We seek well-controlled methods for combining the strengths of each approach (while mitigating the weaknesses?)

- New approaches to ICs (see O.Hahn's talk)
- New ways of exploring the response surface
- New ways of modeling bias
- New ways of controlling sample variance
- New ways of emulating components

# Example: modeling bias

- Different "tracers" of large-scale structure are related to the underlying perturbations in density and potential differently.
- The connection between how the tracer clusters and how the matter clusters is known as bias, and dealing with bias is one of the big challenges in modeling large-scale structure.
  - For example, more luminous galaxies tend to be more clustered than less luminous galaxies, even though both trace the same underlying density field.

How do the different 'camps' typically approach modeling the connection between observable and dark matter when fitting data?

# N-body approach: halo model

Most common approach for surveys is the halo model:

Recent review: Asgari, Mead & Heymans, arXiv:2303.08752

- The objects of interest (e.g. galaxies) live in halos.
  - Likely true for galaxies, but HI, kSZ, ... ?
- Occupancy depends only upon current properties.
  - plus halo accretion history, mergers, ... ?
- Occupancy depends only upon halo mass.
  - plus spin, concentration, environment, ... ('assembly bias')
  - preferred mass definition ($M_{200b}$, $M_{200c}$, $M_{\rm vir}$)?
- Galaxies can be broken up into centrals and satellites (fly-throughs?).
- The full "halo occupation distribution" [HOD: $P(N_{\rm gal}|{\rm halo})$], can be derived from the mean e.g. $\langle N\rangle(M_h)$.
- Radial and velocity profiles are universal/known.

# N-body approach: halo model

- Run a simulation and identify halos
  - ... and possibly subhalos
- Use HOD to populate halos with mock galaxies
- Perform **any** analysis you like on these galaxies, including measuring their power spectra.
- Can also forward model observational effects like fiber assignment, imaging systematics, complex selections.
  - ... if you trust the model you put into the simulation for each of these, e.g. small-scale clustering for fiber assignment in redshift surveys or photo-$z$-IA interaction in lensing.
- As for PT, combine this with ‘emulators’ or ‘ML’.

# PT approach: (symmetries based) bias expansion

Standard in PT is the 'bias expansion':

- Write $\delta_{\rm gal}$ as a functional of the initial (long wavelength) density, velocity and potential fields: $\delta_{\rm gal}[\delta, \partial \mathbf{v}, \partial\partial\Phi, \cdots]$
- Can't compute this functional, so expand it ...
- Coefficients of an expansion in e.g. $\delta$ are bias coefficients (e.g. J.Stadler's talk!):

$$\delta_{\rm gal}(\mathbf{x}) = b_1 \delta_m(\mathbf{x}) + b_2 \delta_m^2(\mathbf{x}) + \cdots + \text{stochastic} + \cdots$$

- Bias coefficients incorporate our uncertainty about complicated galaxy formation physics in addition to UV effects (automatically includes "assembly bias").
  - Dark matter halo formation, merger history, ...
  - Chemistry and gas cooling.
  - Star formation, SNe, AGN
  - Thermal and kinetic feedback
  - Background radiation

# Bias and EFT

- Similar to "EFT" philosophy: keep all operators obeying the symmetries in an expansion in derivatives.
- While the process that form and shape galaxies and other objects are complex, all such objects arise from simple initial conditions acted upon by physical laws which obey well-known symmetries.
- For non-relativistic tracers these are
  - the equivalence principle
  - translational, rotational and
  - Galilean invariance.
- This highly restricts the kinds of terms that can arise in a bias expansion, no matter how complex the history. Often fewer parameters than in HOD models!

Symmetry arguments are extremely powerful for bias since we really don't understand the small-scale physics of bias.

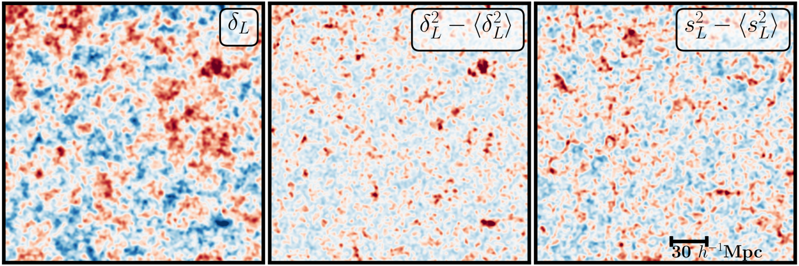
# Simulations and Symmetries

- We can simulate structure formation in a DM-only Universe pretty well.
  - It's the baryonic component that is "hard"!
- Symmetries-based thinking is very powerful.
- Both groups are trying to solve the same problems ...
- Can we have the best of both worlds?
- Use dynamics from N-body simulations, but the "galaxies" (symmetries-based bias technique) from perturbation theory
  - Modi+20, Kokron+21, Hadzhiyska+21, Zennaro+21, Arico+21, Banerjee+22, Zennaro+22, Pellejero+22, Maion+22, Wu+23, Pellejero+23, DeRose+23,...

Hybrid effective field theory (HEFT)

# The hybrid EFT procedure in pictures

Generate initial conditions as per usual ... from $\delta_L$ you can also compute $\delta_L^2$ and the shear field, $s_{ij}$:

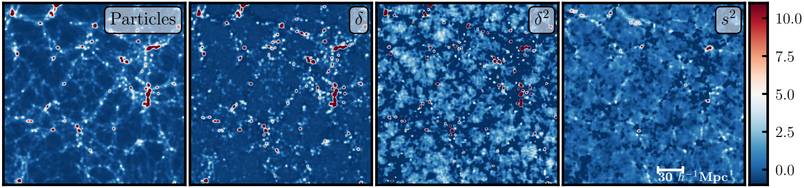



Each particle is assigned the $\delta_L$, ... at its initial position.

Kokron+21

# The hybrid EFT procedure in pictures

Advect the particles to their final positions using the full N-body dynamics (i.e. run the simulation), and bin using weights 1, $\delta_L$, $\delta_L^2$, etc.

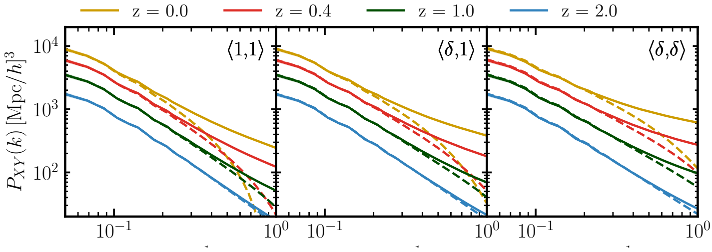



(No need for halo or subhalo finding, merger trees, etc.)

$$\delta_{\rm gal}(\mathbf{x}) = b_1 \delta_m(\mathbf{x}) + \cdots \quad \Rightarrow \quad P_{\rm gal}(k) = b_1^2 P_{mm} + \cdots$$

# The hybrid EFT procedure in pictures

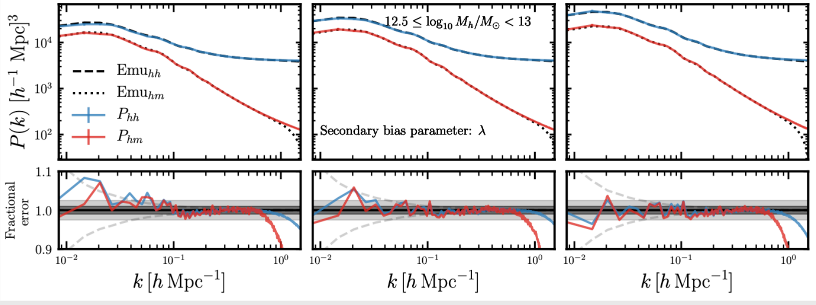
Take all of the cross-spectra, $P_{XY}(k)$ using standard FFT methods, e.g.




The power spectrum for any biased tracer, or the cross-spectrum between any two tracers, is a linear combination of these "basis spectra" (10 in all) with analytic "bias dependence": $\sum_{ij} b_i b_j P_{ij}$.

# Can push into the non-linear regime

Can fit mock catalog data for "3 $\times$ 2pt analyses" to 1-2% even for samples with assembly bias and other complex selections and even including hydrodynamics/"baryons".




Kokron+21

# But now we need an emulator

- Several investigations have now shown that this approach works, and works well.
  - Handles auto- and cross-correlations of even complex selections (e.g. colors or SFR, ...).
  - An additional advantage is that you are now using the same parameters (and language) as the PT model(s) that are routinely used to model BAO and RSD in 3D surveys! Facilitates 'joint' analyses.
- But our 'theory' is now 'simulation' and so we need to know the component spectra as a function of cosmology.
- Can't run a simulation for every cosmology we wish to test!
- But the dependence on parameters is "smooth", so we can interpolate/emulate the spectra.
  - No harder than what we already do for $P_{mm}$.

# Aemulus-$\nu$

- New emulator for galaxy-galaxy, galaxy-matter and matter-matter clustering.
- $150 \times 1\, h^{-1}$Gpc N-body simulations in $w\nu\Lambda$CDM parameter space.
  - Neutrinos handled as a second particle species with PM forces
  - Span wide range of $\sigma_8$
  - $< 1\%$ $P_{mm}$ for $k < 1\, h\,\mathrm{Mpc}^{-1}$ over $0 \leq z < 3$
- To pull this off we've made use of a number of PT-dependent 'tricks' in addition to the bias expansion ...

# Control Variates

- Simulations always have limited dynamic range
- In particular large scales are often "noisy" due to sample variance (from the particular realization of the ICs).
  - Especially true for simulations of high resolution, or including hydrodynamics, or RT, where boxes tend to be 'small'.
- These large scales contain very important cosmological information that we want to get 'right', but ...
- If we are running grids of models, don't want to have to run many realizations for each cosmology to average down this scatter.
- PT works very well on large scales!
  - We shouldn't need to simulate linear theory!!!
- Use control variates to reduce sample variance ...

# Background: Control Variates

(First introduced into LSS by Chartier & Wandelt as "CARPool")

- Imagine I want $\langle \mathbf{x} \rangle$ but realizations of $\mathbf{x}$ are expensive to produce.
  - Example, the matter power spectrum or halo power spectrum.
- Further assume I can cheaply produce $\mathbf{c}$, where $\mathbf{c}$ is correlated with $\mathbf{x}$ and $\mu_c = \langle \mathbf{c} \rangle$ is known.
  - $\mathbf{c}$ is known as the *control variate*.
  - Example: $\mathbf{c}$ is the density power spectrum in the Zeldovich approximation (lowest order LPT).
  - Note: we need to <u>know</u> $\mu_c$ but we don't require $\langle \mathbf{c} \rangle = \langle \mathbf{x} \rangle$
- If we form

$$\mathbf{y} \equiv \mathbf{x} - \beta\left(\mathbf{c} - \mu_c\right)$$

then $\langle \mathbf{y} \rangle$ is an unbiased estimator of $\langle \mathbf{x} \rangle$ for any $\beta$.

# Background: Control Variates

$$\mathbf{y} \equiv \mathbf{x} - \beta\left(\mathbf{c} - \mu_c\right) \quad \Rightarrow \quad \langle \mathbf{y} \rangle = \langle \mathbf{x} \rangle$$

- Now choose

$$\beta^{\star} = \frac{\mathrm{Cov}[\mathbf{x}, \mathbf{c}]}{\sigma_c^2}$$

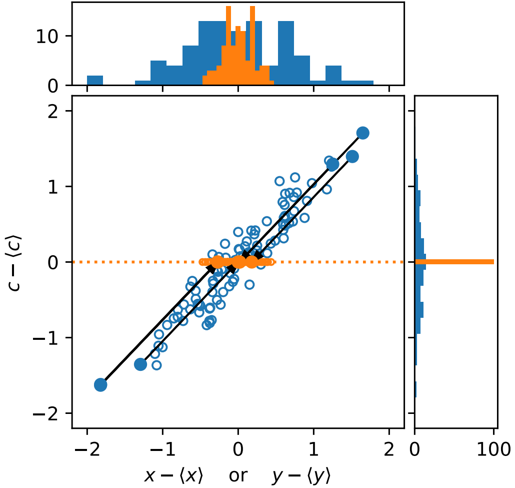
  (really a matrix expression but frequently just approximate as diagonal).
- Then

$$\mathrm{Var}[\mathbf{y}] = \mathrm{Var}[\mathbf{x}]\left(1 - \rho_{xc}^2\right) \quad , \quad \rho_{xc} \equiv \frac{\mathrm{Cov}[\mathbf{x}, \mathbf{c}]}{\mathrm{Std}[\mathbf{x}]\mathrm{Std}[\mathbf{c}]}$$

- If $\rho_{xc} \approx 1$ then **y** is a very low noise/scatter quantity that well-estimates $\langle \mathbf{x} \rangle$.
- Heuristically if **c** fluctuates above $\mu_c$ then **x** probably also fluctuated "up" so you should correct it down.

# Control Variates

Can visualize CV as shifting points along their degeneracy direction to $\mathbf{c} = \langle \mathbf{c} \rangle$. This tightens the distribution and makes it easier to estimate the mean.

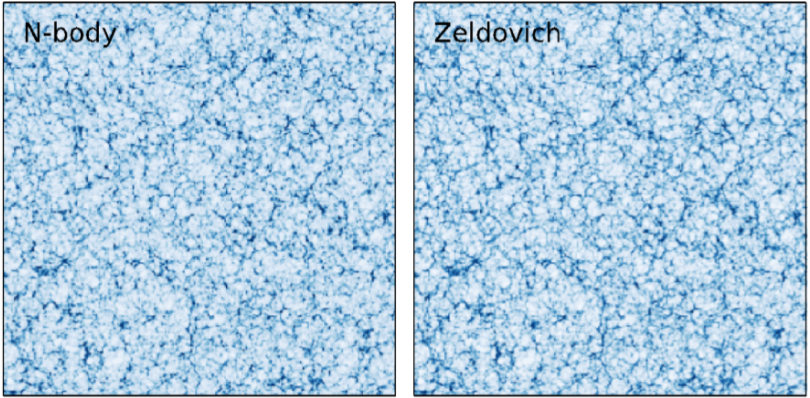


# Control Variates

- Need a control variate that is well correlated with final density field.
- Want **c** to have a known mean – any error in $\mu_c$ leads to bias in final result!
  - Can estimate $\mu_c$ by Monte-Carlo, but expensive.
  - Even if surrogate is 100× faster than full simulation getting to 10% relative error on $\mu_c$ doubles your CPU time!
  - DESI team used 24M CPU hours and 400TB of storage running `FastPM` simulations to estimate $\langle P \rangle$.
  - Paying $N_{\mathrm{sim}}^{-1/2}$ price is bad when demanding high accuracy (e.g. 1% requires $10^4$ 'fast' sims).
  - What about ML? If can train something to correlate, then could do mean via Monte-Carlo.
- Want to be able to include bias and redshift-space distortions.

# Zeldovich and the cosmic web!

We have known for many decades that the Zeldovich approximation ($1^{\mathrm{st}}$ order LPT) ‘predicts’ the cosmic web with quite high fidelity (i.e. $\rho \approx 1$ on large scales):

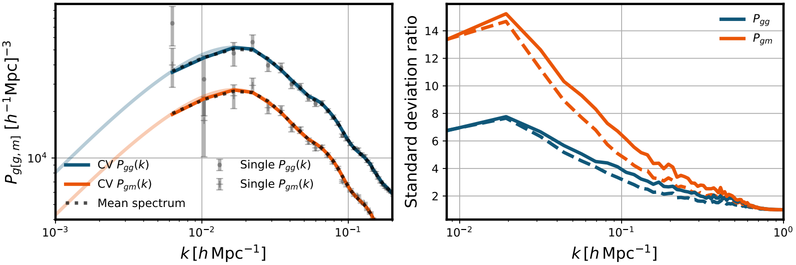


# Zeldovich Control Variates

- Predictions for the Zeldovich mean can be computed analytically
  - Corresponds to “free field theory” so don't even need the simplifications we normally employ for higher-order PT.
  - With some cleverness can be done very efficiently with FFTs.
  - Full run time is $<$ minutes for analytical calculation and 10s of CPU hours for $P(\mathbf{k})$ of the Zeldovich ICs (c.f. 24M CPU hours!).
- If you start your simulation using Lagrangian PT (e.g. the Zeldovich approximation, or higher order) then you already have Zeldovich field.
- No need to generate special ICs, rerun your simulation or do anything ‘fancy’!!
- This technique is mathematically rigorous, and works for all spectra (including higher-order spectra, etc.) unlike e.g. paired and fixed method.

# Accurate predictions from small boxes!

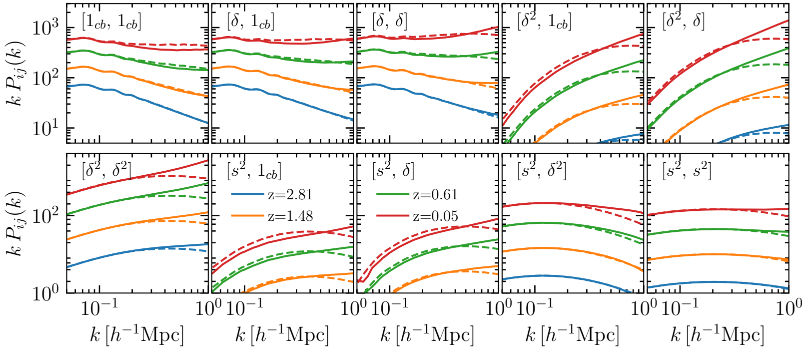
An example of measuring $P_{gg}$ and $P_{gm}$ from a single $1\,h^{-1}$Gpc box, after applying CV and compared to the average of 100 such boxes (light lines extend to low $k$ using PT).
[Even better performance for $P_{mm}$; not shown to reduce clutter.]




Kokron+22

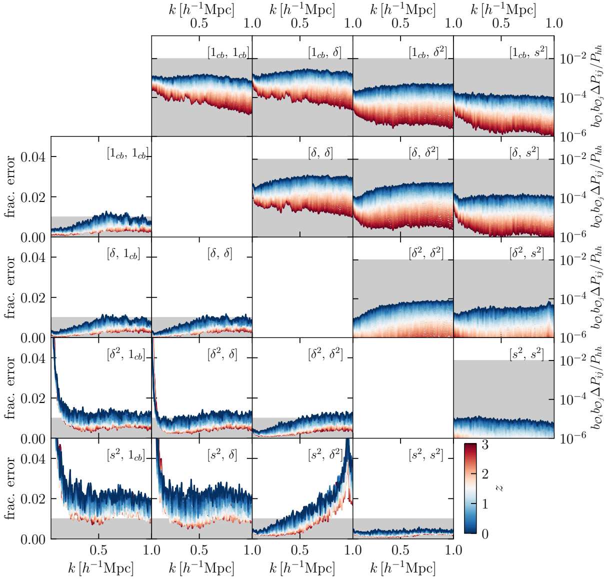
# Almost noiseless predictions!

When we combine with perturbation theory at ultra-large scales the result is (almost) noiseless predictions for the component spectra:




DeRose+23

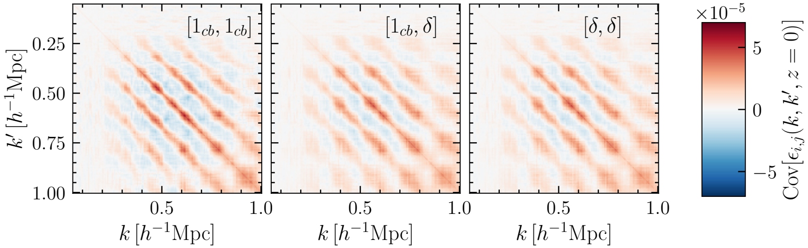
# Interpolation "error"




DeRose+23

# Including theory "error"

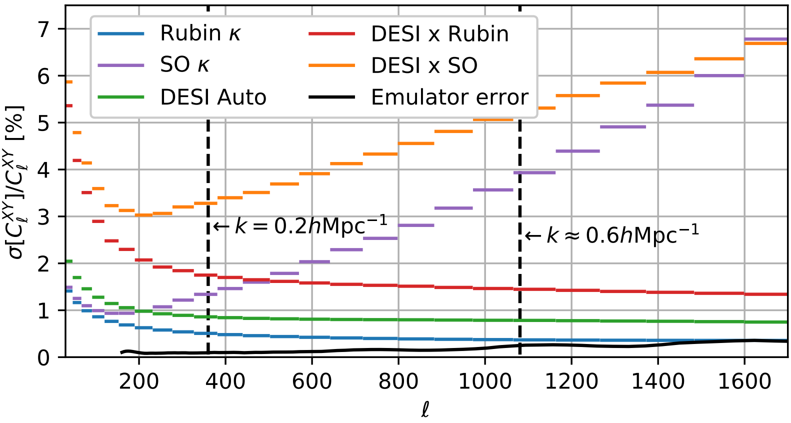
We also provide a covariance matrix of the interpolator error that can be included when performing fits!




DeRose+23

# Forecasted performance

The emulator error (black line) should be much smaller than the statistical errors of next-generation surveys:




DeRose+23

# CV are very flexible

We have used control variates for

- Component spectra, $P_{XY}(k)$, for our emulator.
- Redshift-space power spectrum and correlation function multipoles (actively used in DESI).
- Post-reconstruction $P_\ell$ and $\xi_\ell$.

but the technique is super powerful.

- Not restricted to two-point functions – can be covariance or higher order functions, mass functions, ...
- Not restricted to N-body simulations – can be noise bias terms in survey mocks, instrument simulations, ...

You only need a correlated variable for which the mean can be reliably estimated!

# Conclusions

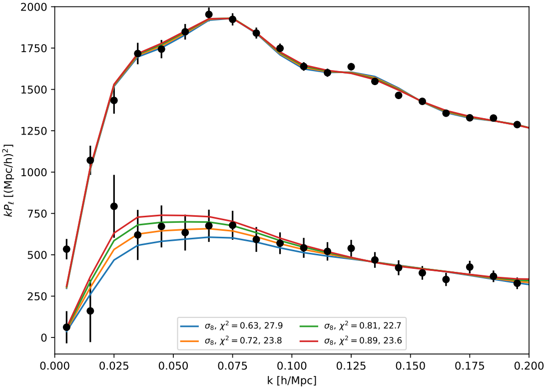
We are in the midst of the "golden age of cosmological surveys".

- Desire for models capable of handling intra- and inter-survey analysis.
- Increasing survey power is driving a renaissance in analytic models of large-scale structure.
- Combinations of theory, ML and simulations increasingly used.
- These approaches can work very well together, in some cases leading to the 'best of both worlds'.
- We've combined many of these ideas to produce a modeling framework capable of handling data from many different surveys in a robust and theoretically well-controlled manner.
- Applications to various datasets are "in progress".

*The End!*

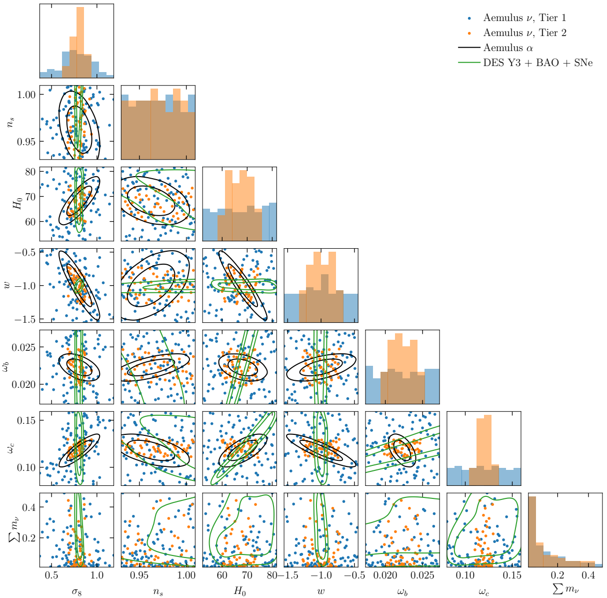
# The PT view of data

Constraining power comes from large scales ... but small scales help constrain transition to non-linearity.




Chen+22

# Emulator design: $w\nu\Lambda$CDM

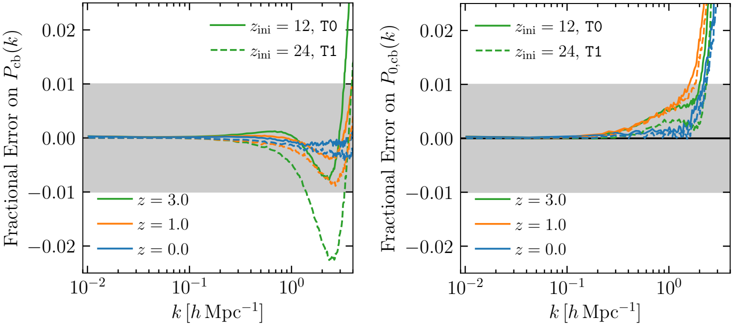



DeRose+23

# Initial conditions

Simulations are squeezed by needing to control errors from a decaying mode and particle discreteness. Extremely challenging if high accuracy is needed.

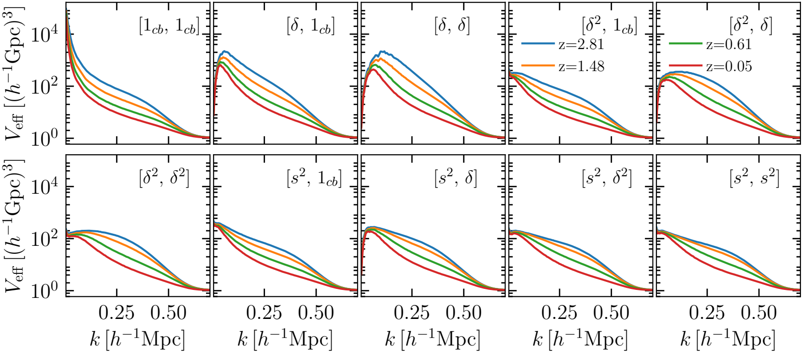
Starting as late as possible with high order LPT is the best strategy!




DeRose+23 following Michaux+21

# Enormous 'effective' volume at modest cost!

While we run only a single 1 $h^{-1}$Gpc box at each cosmology, the sample variance is as if we had run hundreds or thousands ...

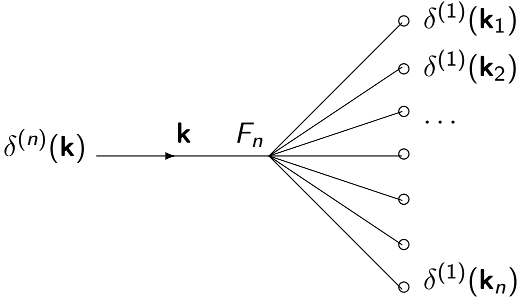



DeRose+23

# Eulerian PT

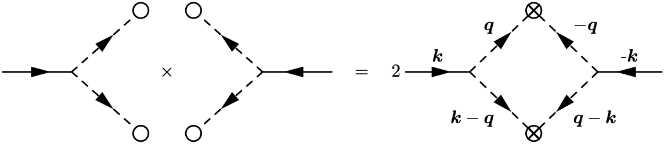
Expand $\delta(\mathbf{k})$ as a power series in the linear solution:

$$\delta^{(n)}(\mathbf{k}) = \int \prod_{i=1}^{n} \frac{d^3k_i}{(2\pi)^3}(2\pi)^3\delta^{(D)}\left(\sum \mathbf{k}_i - \mathbf{k}\right) F_n\left(\{\mathbf{k}_i\}\right)\delta(\mathbf{k}_1)\cdots\delta(\mathbf{k}_n)$$



# Eulerian PT

Taking the expectation value "joins" pairs of $\delta^{(1)}$ together to form a power spectrum. So the propagator at 1-loop contains a contribution like e.g.:

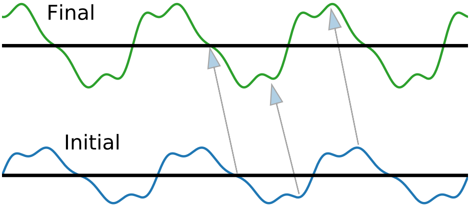



$$= 2\int \frac{d^3q}{(2\pi)^3} F_2(\boldsymbol{q}, \boldsymbol{k}-\boldsymbol{q}) F_2(-\boldsymbol{q}, \boldsymbol{q}-\boldsymbol{k}) P_0(q) P_0(|\boldsymbol{k}-\boldsymbol{q}|)$$

Many other rules look similar to particle or condensed matter physics – use path integrals, generating functions, cutoffs, EFT, RG flows, etc.

# IR resummation

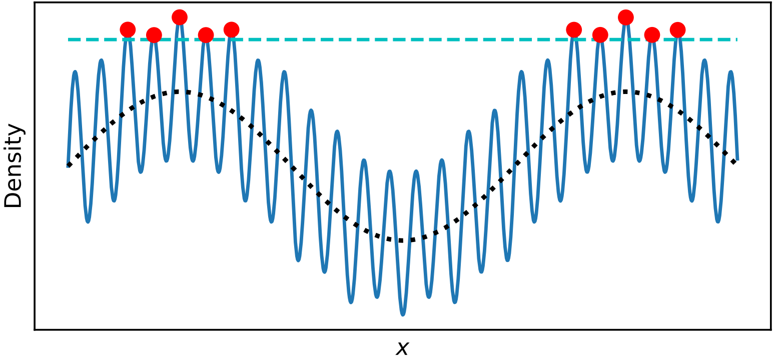
- In addition to problems in the UV, there are issues in the IR.
- A lot of the difference between $\delta^{(\mathrm{non-lin})}(\mathbf{x})$ and $\delta^{(1)}(\mathbf{x})$ comes from displacement (advection).




- The displacement is driven by large-scale tidal fields.
- In "standard" perturbation theory this effect converges slowly.
- Need to "resum" the long-wavelength displacements
  - The Lagrangian formulation of PT is ideally suited to understanding "IR resummation".
  - Impacts topics like "reconstruction", primordial features, relative velocity effect, ...

# Bias, peaks and EFT

- To make contact with galaxies, QSOs, 21 cm, Ly$\alpha$, etc. we need to include bias.
- Simplest (toy) model: galaxies form at peaks in the initial density field:



# Massive neutrinos

- Galaxies probe the $c + b$ field while lensing probes the matter.
- At linear level use $P_{cb}(k)$ for galaxies and $P_{cb,m}$ for galaxy-lensing cross-correlation.
  - Good to sub-percent level (e.g. Bayer+21)
- If care is taken with normal ordered bias operators, can use $P_{cb,m}$ in loops with corrections of order $f_\nu P_{\rm lin}^2 \ll 1$ and be correct even in the "transition regime" from clustered to free-streaming neutrinos.

# Aside: PT and ML

- These PT models are fast to compute, but 'fast' means seconds not milli-seconds.
  - Running CAMB or CLASS takes $1 - 10\,\mathrm{s}$ depending on settings.
- ML can be used to generate emulators of the linear theory and PT predictions which are then **very** fast to evaluate.
  - Pay price once "up front" and future evaluations are very cheap.
- Can be used to generate enormous training sets, so this is an "easy" ML problem.
  - Taylor series, polynomial chaos expansions, neural networks with different architectures, ...
- Since PT contains much of the parameter response, can be used in planning more complex campaigns – or comparing ML methods.