OPRE Innovative Methods Meeting
Washington, DC
September 22, 2016

Designing Controlled Trials with the Power of Optimization

Nathan Kallus

Asst Prof @ Cornell University and Cornell Tech

All code is available at www.nathankallus.com

Based on papers:

K. Optimal A Priori Balance in the Design of Controlled Experiments. To appear in *Journal of the Royal Statistical Society: Series B*.

Bertsimas, Johnson, K. The Power of Optimization Over Randomization in Designing Experiments Involving Small Samples. *Operations Research*.

Example: Effect of job training

$$Y_{ik} = \text{potential outcome of subject } i \text{ if given treatment } k$$

$$\text{SATE}_{kk'} = \frac{1}{n}\sum_{i=1}^{n}(Y_{ik} - Y_{ik'}),\ \text{PATE}_{kk'} = \mathbb{E}\,\text{SATE}_{kk'}$$

$$\hat{\tau} = \text{difference of group mean outcomes}$$

$W_i = 1$	$W_i = 2$
$Y_{i1} = \$20$ thousand	$Y_{i2} = \$12$ million
$Y_{i1} = \$5$ million	$Y_{i2} = \$23$ thousand
$Y_{i1} = \$10$ million	$Y_{i2} = \$500$
$Y_{i1} = \$20$ million	$Y_{i2} = \$16$ thousand

Treatment

Control

Proposals for a priori balance

(i.e., balance in baseline covariates *X* before randomization and treatment)

- Blocking (Fisher 1935):
  imbalance = –(# perfect exact matches)
- Pairwise-matching (Greevy et al 2004 “optimal”):
  imbalance = sum of pair distances
- Re-randomization (Morgan and Rubin 2012):
  imbalance = group-wise Mahalanobis metric

- *What is “balance” anyway?*
- *If we know, can we make it **better**?*

What can't balance achieve?

Thm (K '15): If $Y_{i1} + Y_{i2}$ is mean-independent of X_i, then *all* a priori designs yield the same variance.

Generalizes no efficiency loss of blocking irrelevant covariates (Cochran & Cox, 1957) or pairwise matching on irrelevant covariates (Chase, 1968).

No Free Lunch Theorem (K '15) (informally stated)
Without imposing any structure, one cannot get better variance than complete randomization (aka, no balance) using a priori balancing.

Conclusion: balance goes hand in hand with structure
And: if have structure, why not optimize?

What can balance achieve?

- If we balance (or not) a priori then $\mathbb{E}[\hat{\tau} - \mathrm{SATT} | X, Y] = 0$, and $\mathrm{Var}(\hat{\tau}) = \mathrm{Var}(\mathbb{E}[\hat{\tau}|X]) + \mathbb{E}[\mathrm{Var}(\hat{\tau}|X)]$

  unaffected by balancing

- Hope is low $\mathrm{Var}(\mathbb{E}[\hat{\tau}|X]) = \mathbb{E}[B^2(W, \hat{f})|X]$ where $B(W, f) = \frac{2}{n}\sum_{i=1}^{n}(-1)^{1+W_i} f(X_i)$

$$\hat{f} = \mathbb{E}\left[\frac{Y_{i1} + Y_{i2}}{2}\middle| X_i = x\right]$$

- Best case scenario:

$$\frac{\mathrm{Var}(\hat{\tau})}{\mathrm{Var}(\hat{\tau}^{\mathrm{CR}})} \geq 1 - R^2, \text{where } R^2 = \frac{\mathrm{Var}\left(\hat{f}(X_i)\right)}{\mathrm{Var}\left(\frac{Y_{i1} + Y_{i2}}{2}\right)}$$

Optimal Balance

- Efficient design minimizes $\mathbb{E}[B^2(W, \hat{f})|X]$
- But $\hat{f}$ unknown. Take minimax (or, Bayesian) approach. How to define magnitude (prior)?

- **Pure-strategy optimal design (PSOD):** Draw assignment W uniformly at random from the set of optimizers

$$W \in \arg\min_{W \in \mathcal{W}} \left\{ M_{\mathrm{p}}^2(W) := \max_{f \in \mathcal{F}} \frac{B^2(W, f)}{||f||^2} \right\}$$

Existing designs as optimal

Theorem (K ‘15): Let

$$||f|| = ||f||_\infty = \sup_x |f(x)|$$

then the PSOD is ***incomplete blocking***.

Theorem (K ‘15): Let

$$||f|| = ||f||_{\mathrm{lip}} = \sup_{x \neq x'} \frac{|f(x) - f(x')|}{d(x, x')}$$

then the PSOD is ***optimal pairwise matching***.

Existing designs as optimal

- Problem: convergence is too slow*!*
  - As dimension grows even modestly
    - number of factors to stratify on grows
    - neighbors become farther apart
  - This structure can be too loose
    ⇒ not enough power with small samples
- $\mathrm{Var}(\mathbb{E}[\hat{\tau}|X])$ has *logarithmic* convergence $(1/n^c)$ for
  - Pairwise matching
  - Well-specified linear parametric form but balance not *fully optimized* (e.g. re-randomization)
  - Misspecified form (e.g. coarsening)
- What if we can achieve *linear* convergence $(1/2^n)$?

Kernel Matching

- $\mathcal{F} = \overline{\mathrm{span}}(\{\mathcal{K}(x,\cdot)\}_x)$ for PSD kernel $\mathcal{K}(x,x')$
  - Polynomial kernel $\mathcal{K}_s(x,x') = (1 + x^T x'/s)^s$
  - Exponential kernel $\mathcal{K}(x,x') = e^{x^T x\prime}$
  - Gaussian kernel $\mathcal{K}_s(x,x') = e^{-\|x-x'\|_2^2/s^2}$
- (Apply after normalizing or fit rotation using marginal likelihood)
- The PSOD with two treatments is

$$\min_{u\in\{-1,1\}^n \,:\, \sum_i u_i=0} u^T K u \quad \text{where } K_{ij} = \mathcal{K}(X_i, X_j)$$

- Solve using integer optimization solver Gurobi

(free license for academic use; open source alternatives exist)

Variance and Convergence

Theorem (K '15): For kernel matching,

$$1 - R^2 \leq \frac{\mathrm{Var}(\hat{\tau})}{\mathrm{Var}(\hat{\tau}^{\mathrm{CR}})} \leq 1 - R^2 + C_1 n \mathbb{E}\left[\min_W M_{\mathrm{P}}^2(W)\right]$$

"Theorem" (K '15): For a finite dimensional kernel,

$$\mathbb{E}\left[\min_W M_{\mathrm{P}}^2(W)\right] = 2^{-\Omega(n)}$$

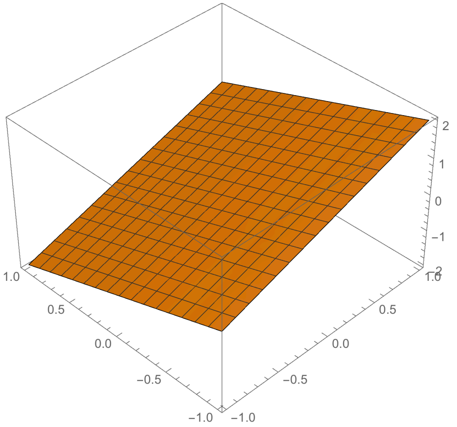
Theorem (K '15): For a universal kernel (e.g. Gaussian or exponential kernel), $\hat{\tau} - \mathrm{SATT} \xrightarrow{\mathbb{P}} 0$.

Asides

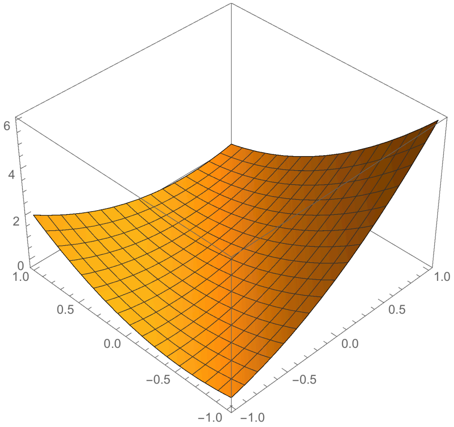
- For finite dimensional spaces, ℓ_2 norm (kernels) versus other norms doesn't matter
  - Instead of quadratic kernel, Bertsimas, Johnson, K method uses ℓ_1 norm to get a more tractable optimization problem (integer-linear program)
- In presence of non-compliance, kernel matching can help accurately estimate causal treatment effects
  - Even if true treatment randomization is not possible (treatment assignment is mere instrument), can still have very accurate measures of effect in small samples

Synthetic examples

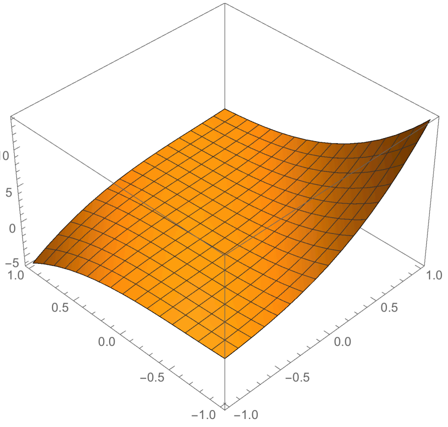
- Constant effect $Y_{i1} - Y_{i2} = \tau$
- $d = 2$ covariates $X_i \sim \text{Unif}\left([-1,\ 1]^2\right)$
- Normalize variance relative to irreducible part
- Fix various $\hat{f}$ functions



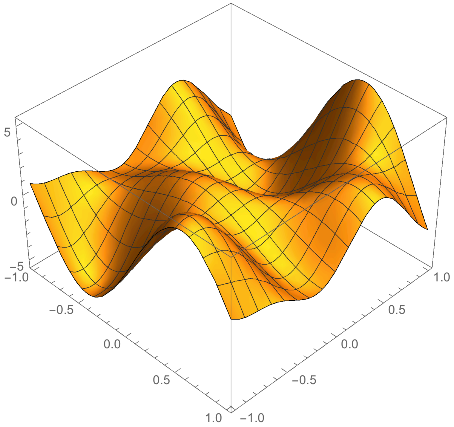
Linear



Quadratic



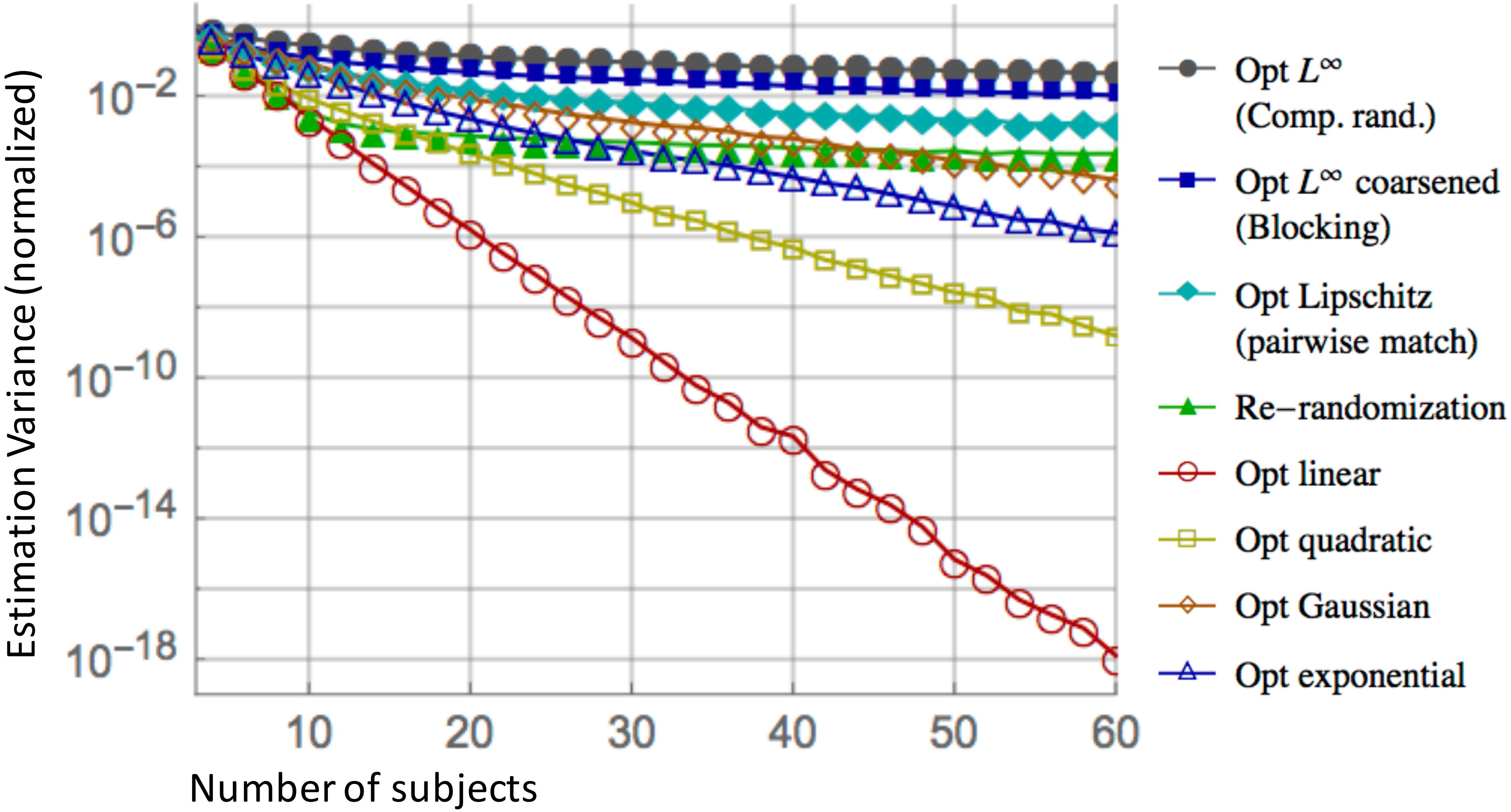
Cubic



Sinusoidal

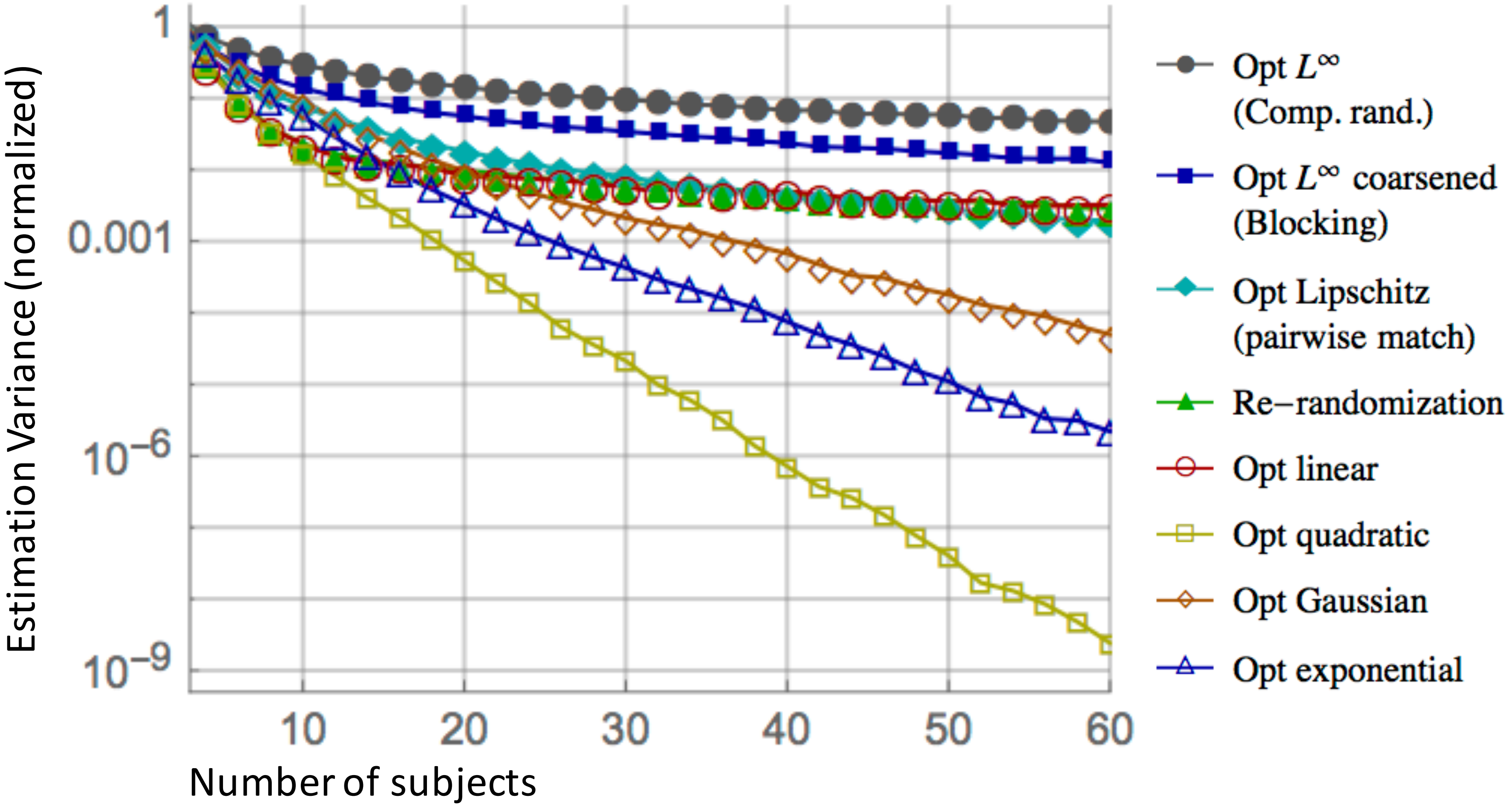
Synthetic examples: Linear effect

$$\hat{f}(x_1, x_2) = x_1 - x_2$$



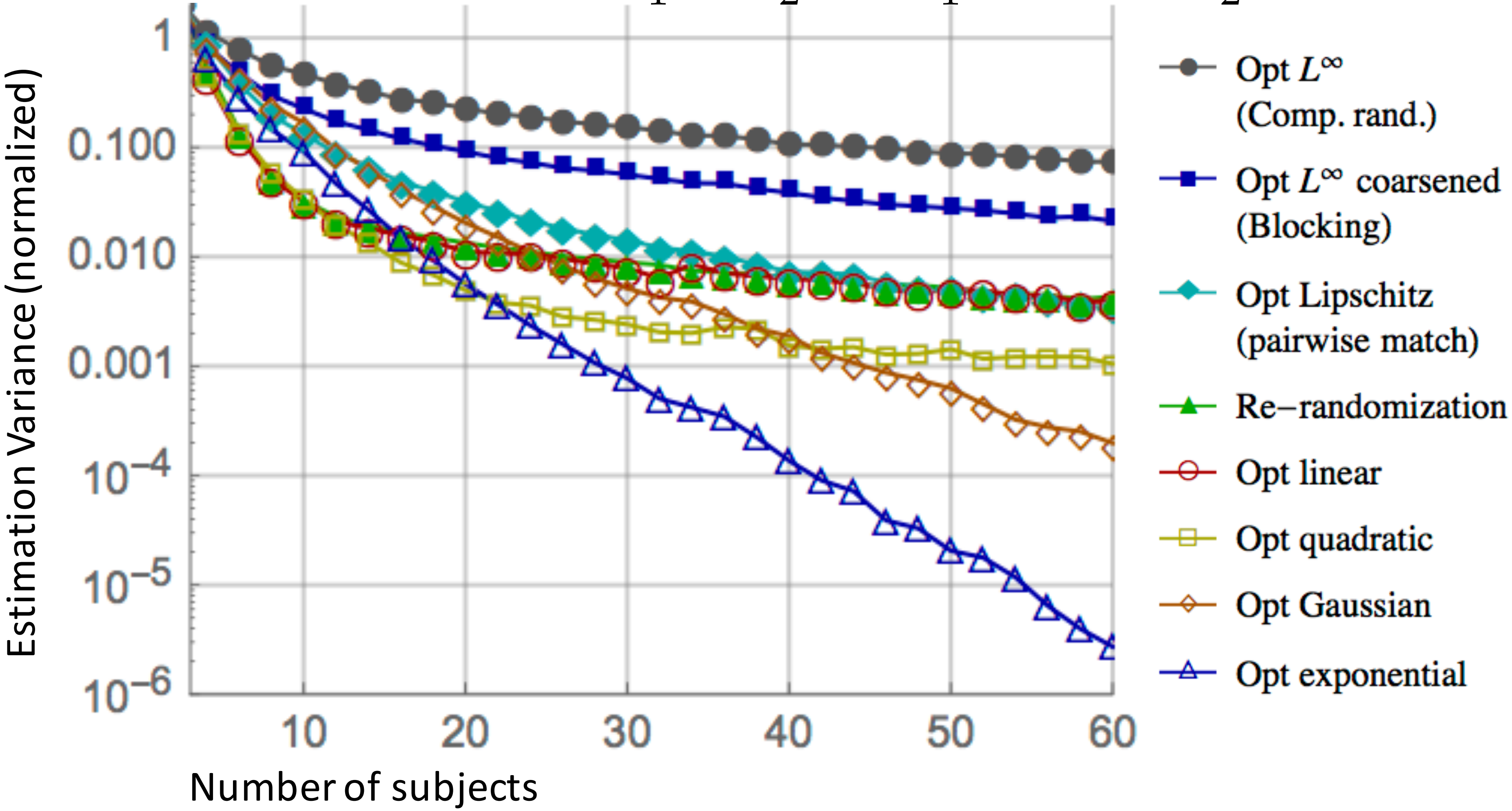
Synthetic examples: Quadratic effect

$$\hat{f}(x_1, x_2) = x_1 - x_2 + x_1^2 + x_2^2 - 2x_1x_2$$



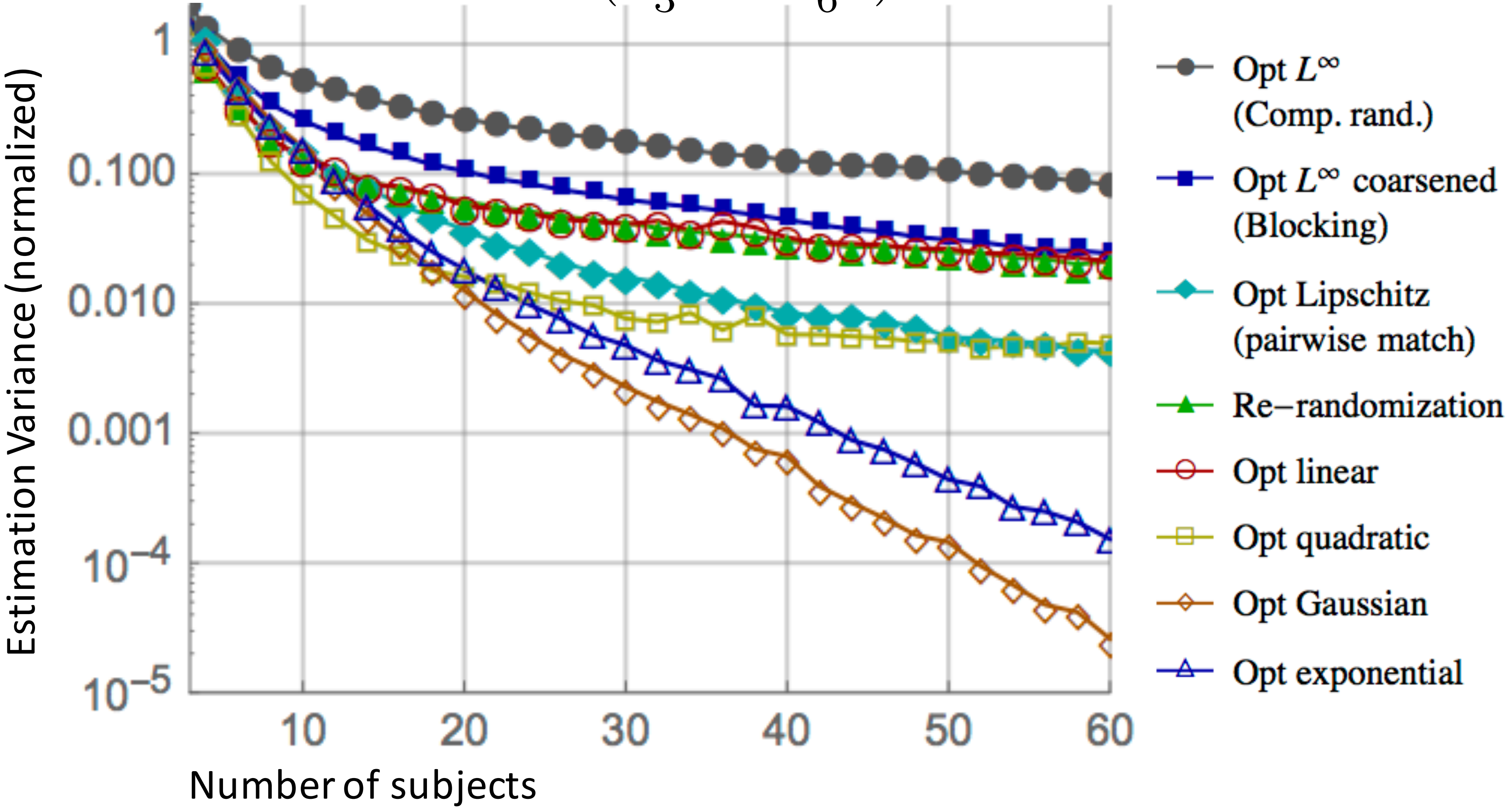
Synthetic examples: Cubic effect

$$\hat{f}(x_1, x_2) = x_1 - x_2 + x_1^2 + x_2^2 - 2x_1x_2 + x_1^3 - x_2^3 - 3x_1^2x_2 + 3x_1x_2^3$$



Synthetic examples: Sinusoidal effect

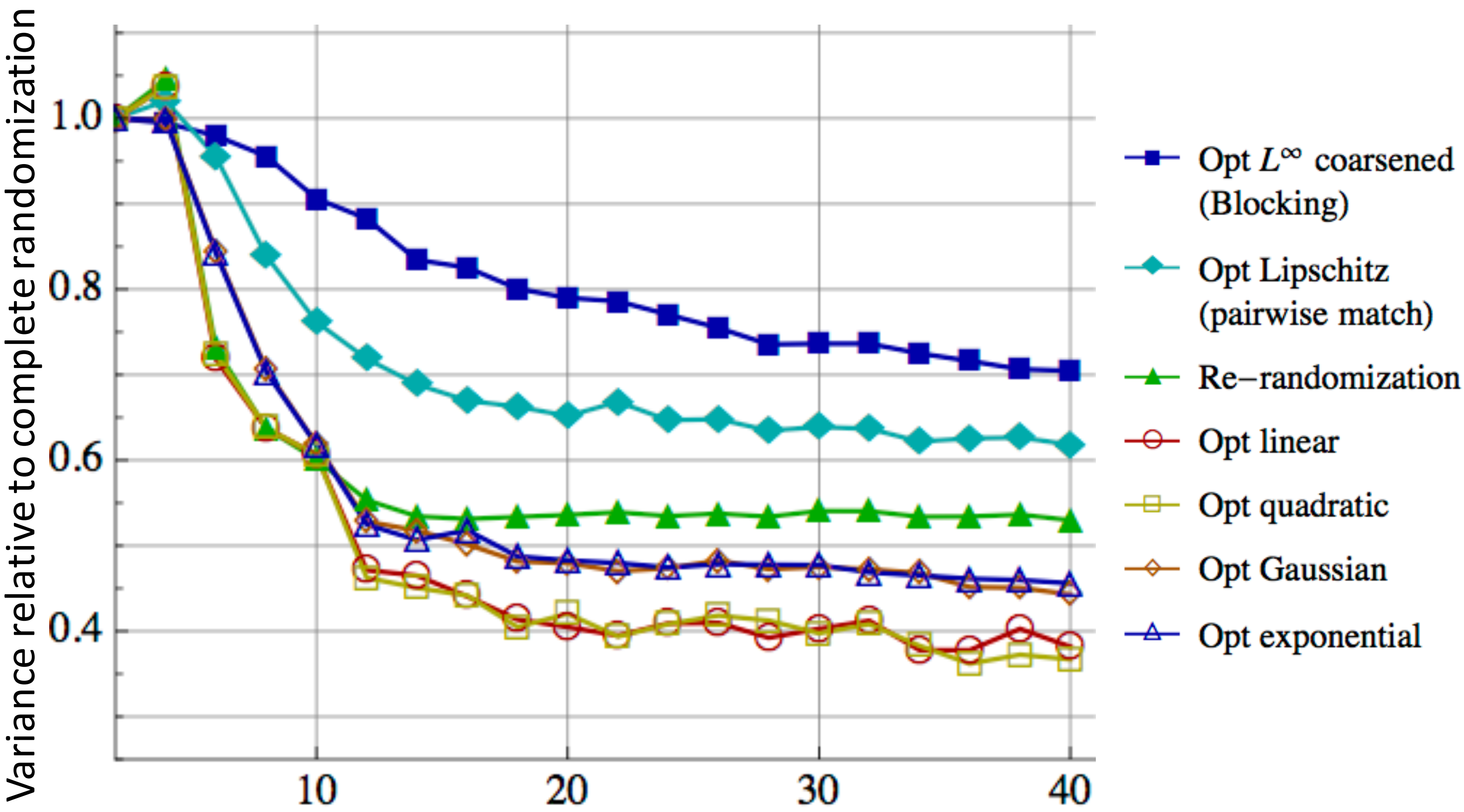
$$\hat{f}(x_1, x_2) = \sin\left(\frac{\pi}{3} + \frac{\pi x_1}{3} - \frac{2\pi x_2}{3}\right) - 6\sin\left(\frac{\pi x_1}{3} + \frac{\pi x_2}{4}\right) + 6\sin\left(\frac{\pi x_1}{3} + \frac{\pi x_2}{6}\right)$$



Experiment with Clinical Data

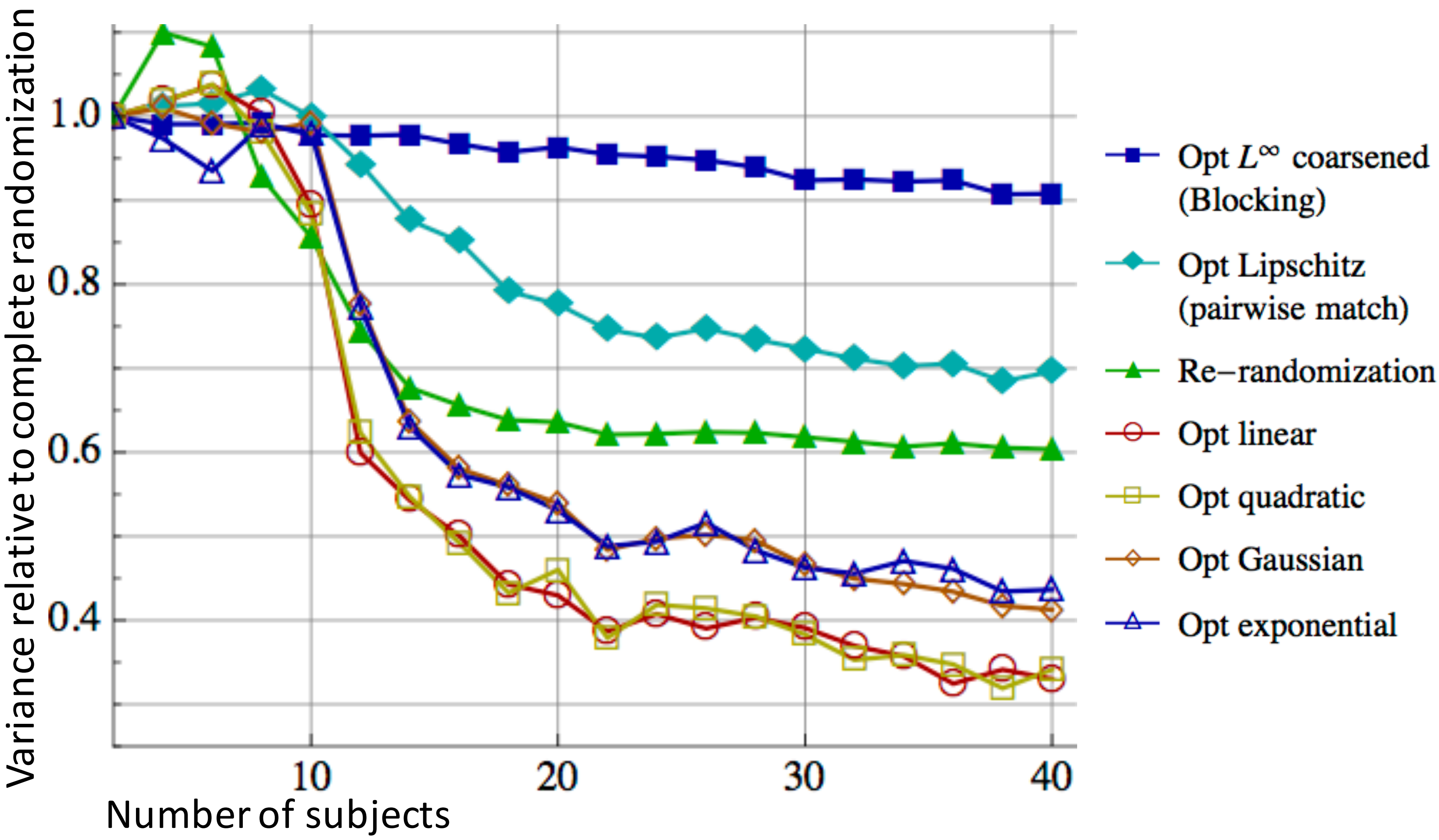
- Diabetes patient dataset:
  - $d = 10$ covariates: age, sex, BMI, blood pressure, blood serum measurements
  - $N = 442$ subjects
  - Y = Change in HbA1C level after one year
- Experiment:
  - What's the effect of a new oral drug on HbA1C?
  - Hidden reality: there's a constant effect
- Either use $d = 4$ most predictive covars selected by post-hoc LARS, or use all $d = 10$ covars

Experiment with Clinical Data, *d*=4



Number of subjects

Experiment with Clinical Data, d=10



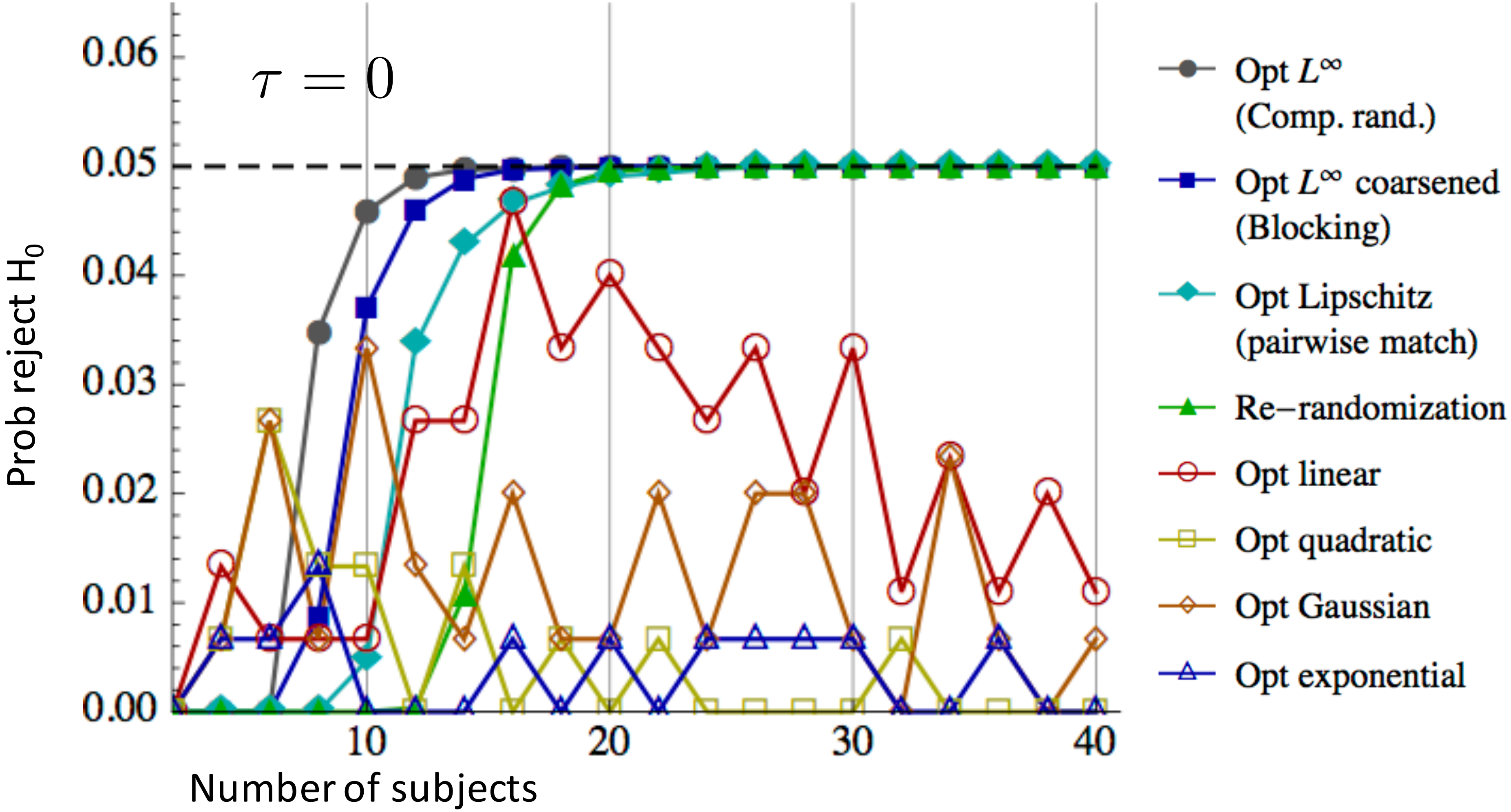
Inference

- If there's not enough randomization (e.g. as in PSOD) then Fisher's randomization test will be underpowered (p = 1).
- Solution: a bootstrap test

1. Draw W^0 from the PSOD for $X_1, \ldots, X_n$, assign, apply treatment, measure outcomes $Y_i = Y_{iW_i^0}$, compute $\hat{\tau}$.
2. For t = $1, \ldots, T$:
    1. Sample $i_j^t \sim \mathrm{Unif}\{1, \ldots, n\}$ iid
    2. Draw W^t from the PSOD for $X_{i_1^t}, \ldots, X_{i_n^t}$
    3. Compute $\tilde{\tau}^t = \frac{1}{p} \sum_{j:W_j^t=1} Y_{i_j^t} - \frac{1}{p} \sum_{j:W_j^t=2} Y_{i_j^t}$
3. The p-value of H_0 is $p = (1 + |\{t \ : \ |\tilde{\tau}^t| \geq |\hat{\tau}|\}|) \,/\, (1 + T)$
   Reject H_0 if $p \leq \alpha$

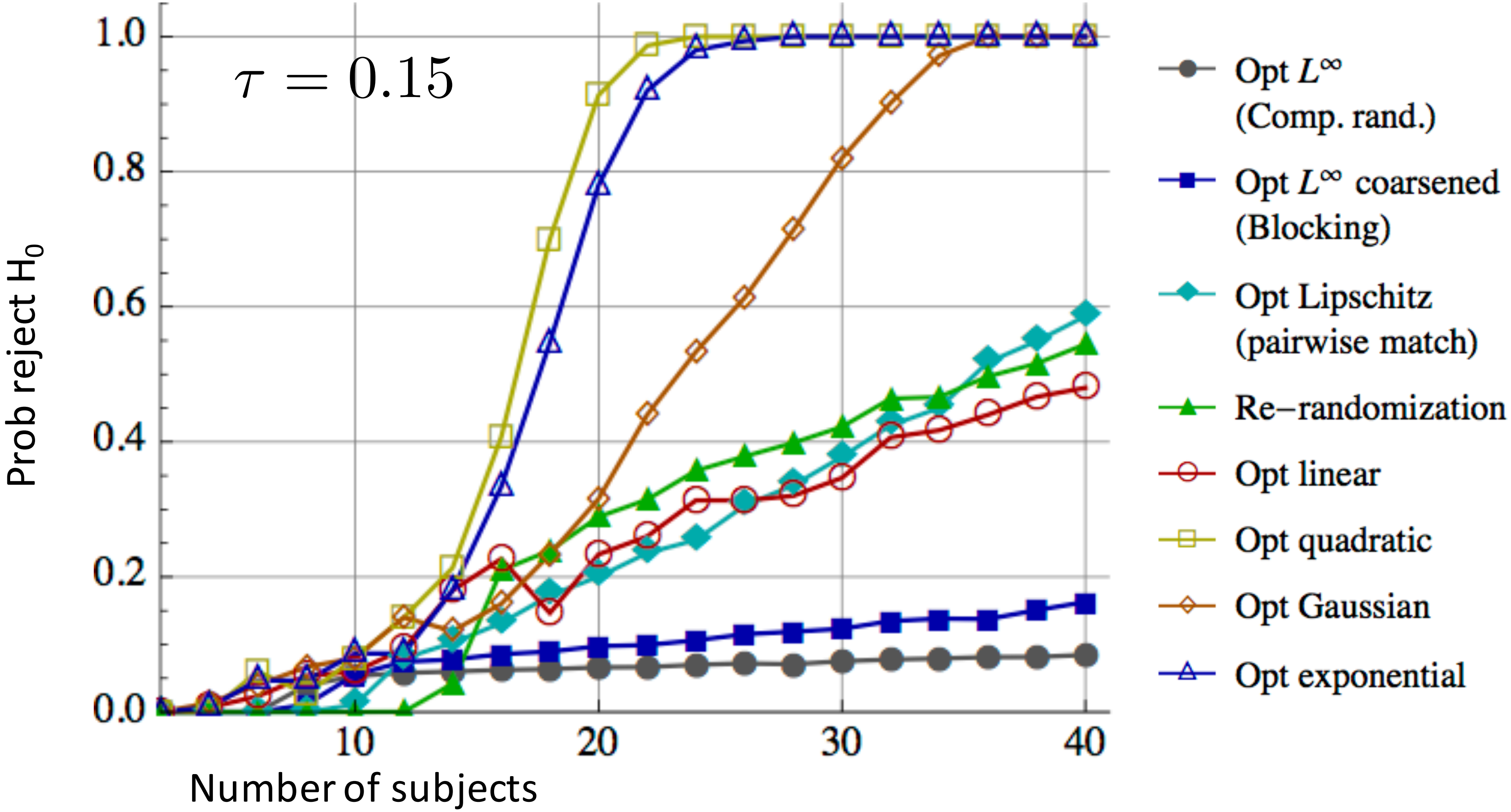
Synthetic examples: Quadratic effect

$$\hat{f}(x_1, x_2) = x_1 - x_2 + x_1^2 + x_2^2 - 2x_1x_2$$



Synthetic examples: Quadratic effect

$$\hat{f}(x_1, x_2) = x_1 - x_2 + x_1^2 + x_2^2 - 2x_1x_2$$



… huh?

- Any notion of balance must go hand in hand with a notion of structure
- This recovers existing balancing designs
- New designs: ***kernel matching***
- Empirical & theoretical evidence of superiority
- Especially important for small samples:

$$2^{-\Omega(n)} \text{ vs } O(1/\sqrt{n})$$

Thank you.

Nathan Kallus

www.nathankallus.com